ROYAL SOCIETY OPEN SCIENCE

rsos.royalsocietypublishing.org

Research

Article submitted to journal

Subject Areas:
Psychology

Keywords:
replication, study selection, consensus

Author for correspondence:
[1] Merle-Marie Pittelkow
e-mail: m.pittelkow@rug.nl

THE ROYAL SOCIETY PUBLISHING

The Process of Replication Target Selection in Psychology: What to Consider?

Merle-Marie Pittelkow[1], Sarahanne M. Field[2], Peder M. Isager[3], Anna E. van 't Veer[4,5], Thomas Anderson[6], Scott N. Cole[7], Tomáš Dominik[8], Roger Giner-Sorolla[9], Sebahat Gok[10], Tom Heyman[4], Marc Jekel[11], Timothy J. Luke[12], David B. Mitchell[13], Rik Peels[14], Rosina Pendrous[15,16], Samuel Sarrazin[17], Jacob M. Schauer[18], Eva Specker[19], Ulrich S. Tran[19], Marek A. Vranka[20], Jelte M. Wicherts[21], Naoto Yoshimura[22,23], Rolf A. Zwaan[24], Don van Ravenzwaaij[1]

[1] Department of Psychometrics and Statistics, Rijksuniversiteit Groningen, Groningen, The Netherlands
[2] Centre for Science and Technology Studies, Leiden University, Leiden, the Netherlands
[3] Oslo New University College, Oslo, Norway
[4] Methodology and Statistics Unit, Institute of Psychology, Leiden University, Leiden, the Netherlands
[5] Leiden Institute for Brain and Cognition, Leiden University, Leiden, the Netherlands
[6] Department of Psychology, University of Toronto, Toronto, Canada
[7] School of Education, Language & Psychology, York St John University, York, UK
[8] Institute for Interdisciplinary Brain and Behavioral Sciences, Chapman University, Orange, USA
[9] School of Psychology, University of Kent, Canterbury, UK
[10] Program in Cognitive Science & Department of Instructional SystemsTechnology, Indiana University, Bloomington, USA
[11] Social Psychology, University of Cologne, Cologne, Germany
[12] Department of Psychology, University of Gothenburg, Gothenburg, Sweden
[13] WellStar College of Health and Human Services , Kennesaw State University, Kennesaw, USA
[14] Philosophy Department and the Faculty of Religion and Theology, Vrije Universiteit, Amsterdam, The Netherlands
[15] Centre for Contextual Behavioural Science, School of Psychology, University of Chester, Chester, UK
[16] Institute of Applied Health Research, University of Birmingham, Birmingham, UK
[17] Maison de santé pluridisciplinaire Pasteur, Chevilly-Larue, France
[18] Department of Preventive Medicine - Division of Biostatistics, Feinberg School of Medicine, Northwestern University, Chicago, USA
[19] Department of Cognition, Emotion, and Methods in Psychology, Faculty of Psychology, University of Vienna, Vienna, Austria
[20] Faculty of Social Sciences, Charles University, Prague, Czech Republic
[21] Department of Methodology and Statistics, Tilburg University, Tilburg, The Netherlands
[22] Research Organization of Open Innovation and Collaboration, Ritsumeikan University, Osaka, Japan
[23] Research Fellow of the Japan Society for the Promotion of Science [24] Department of Psychology, Education, and Child Studies, Erasmus University Rotterdam, Rotterdam, the Netherlands

Increased execution of replication studies contributes to the effort to restore credibility of empirical research. However, a second generation of problems arises: the number of potential replication targets is at a serious mismatch with available resources. Given limited resources, replication target selection should be well-justified, systematic, and transparently communicated. At present the discussion on what to consider when selecting a replication target is limited to theoretical discussion, self-reported justifications, and a few formalized suggestions. In this Registered Report, we proposed a study involving the scientific community to create a list of considerations for consultation when selecting a replication target in psychology. We employed a modified Delphi approach. First, we constructed a preliminary list of considerations. Second, we surveyed psychologists who previously selected a replication target with regards to their considerations. Third, we incorporated the results into the preliminary list of considerations and sent the updated list to a group of individuals knowledgeable about concerns regarding replication target selection. Over the course of several rounds, we established consensus regarding what to consider when selecting a replication target.

The last two decades have brought uncertainty to the empirical sciences. Researchers have grown increasingly sceptical of the reliability of previously accepted findings, a situation characterized as a crisis of confidence, reproducibility, replication, or credibility [1,2]. In psychology, the crisis narrative might have many origins: Ioannidis's controversial article [3], some uncovered scientific fraud cases in the Netherlands [4], the publication of eye-catching findings of extra-sensory perception [5], and a series of methodological papers describing the ease with which results can be covertly pushed into the desired direction [e.g., 5–7]. This narrative has since gained momentum as the centre of fiery debates, has led to a substantial – and growing – body of literature, and has been the catalyst behind the foundation of countless practical initiatives to improve the reliability and quality of empirical, psychological research.

Many in the scientific community have chosen to challenge outdated practices and transform science for the better. While many initiatives aim to dismantle the academic publishing system, or help researchers educate themselves on good scientific practice, other endeavours grapple with problems with the findings themselves. One key element of these efforts involves various forms of replication. Close replications aim to mirror the original study (OS) as closely as possible, allowing for example for better estimates and correction of false positives, whereas conceptual replications change elements of the OS to allow for understanding boundary conditions (e.g., by changing measurement and manipulations) and theory building of a phenomenon. A large increase in articles concerned with theoretical and philosophical discussions on replication and replicability is coupled with a sharp uptick in the number of empirical replication studies being conducted [for numbers until May 2012, see 8]. In psychology, one example is the wide-spread replication attempt by the Open Science Collaboration, which demonstrated that more than half of the empirical findings under scrutiny did not replicate [9]. Only a third of the original studies (36.1%) suggested a statistically significant effect (i.e., $p < .05$) and less than half (41.9%) of the original confidence intervals included the replicated effect size [9].

Increased interest in the discussion and execution of replication studies contributes to the active effort to restore credibility to scientific research, including psychological research. However, it brings with it a second generation of problems. Among these is the fact that the number of potential replication targets is at a serious mismatch with the resources available for replication studies, both in terms of human labour and in terms of available funds. As one example, in a separate project author A.E.v.t.V. and P.M.I aim to replicate original research in social neuroscience [10]. Even restricting their candidate set to studies using fMRI in the last ten years, they currently have a pool of over two thousand potential targets to select from. The rate at which empirical studies in psychology are published has been growing exponentially for the past century. Simultaneously, the rate at which original studies are replicated is very low. The replication rate in social sciences and psychology alike has been estimated at around 1% [8,11], though the rate is difficult to estimate exactly. While the pile of potential replication targets is growing at an exponential rate, funding for replication is developing more slowly. This results in an enormous back-log of non-replicated research to contend with.

To accommodate the need for replication studies, funding opportunities targeting replication studies that have emerged range from broad scale funding opportunities in the Biomedical Sciences [e.g., 12], Social Sciences and Humanities [e.g., 2,9,13], or Educational Sciences [e.g., 14], to specific initiatives calling for replication in pre-specified areas [e.g., 15]. Even so, grants for replications receive many good proposals, but can only fund a low percentage of them. For example, the Dutch funder NWO could only fund around 10% of submitted replication studies [16]. Though there is an increase in the number of funding opportunities, they remain relatively scarce and overall resources for replication studies remain limited.

Another stumbling block in the road toward regaining certainty and credibility through conducting replication studies is the way in which studies are selected as replication targets. As we have argued in recent publications, target selection is haphazard and often poorly motivated

[for instance, because replicating authors doubt the veracity of original authors or their findings; see 17], and does not make the best use of what scarce resources are available [18,19]. Some authors have suggested ways to select replication targets, such as using cost-benefit analysis [20], employing Bayesian decision-making strategies [21], or selecting at random [22]. While at first glance suggestions on how to select replication targets might appear quite different, common themes do exist. In a comprehensive review, Isager and colleagues [16]) identified four factors often considered when deciding what is worth replicating: (1) value/impact, (2) uncertainty, (3) quality, and (4) costs and feasibility.

Whatever the reasons for selecting a particular replication target, we believe that communicating how the eventual decision was reached is very important. At present, there is no consensus as to what characterizes a study "worth replicating" or "in need of replication". Regardless of whether or not consensus on this matter can possibly be achieved, clearly communicating one's reasoning behind selecting a replication target enables others to understand, and evaluate the decision. To spend limited resources for replication studies wisely, it is in the interest of both researchers and funding agencies to replicate studies that make sense and that make good use of the resources. Having a transparent logbook of why targets are selected for replication is a first step towards spending limited resources well.

To be clear, we believe that science would benefit from transparently reporting the decisions that led to the genesis of *all* studies. However, we argue that there is good reason to consider the decision process for replication studies separately from original studies. First, the motivation of and reasoning behind replication studies might differ. While many original studies explore new claims based on theoretical reasoning and previous literature, replication studies have in the past frequently been motivated by the intent to corroborate existing empirical results. Second, the room for a replication to add to a field's knowledge base can be more readily quantified since the primary function of a replication is to reduce uncertainty about existing results (whereas original research can have many different functions, some of which are hard to represent quantitatively). Therefore the selection process may be optimized more easily for replication studies. Third, due to the lack of being able to play the "novelty card" when justifying the study authors may be facilitated by a systematic approach. With replication and self-correction being deemed important elements of a scientific field [23], a more systematic and transparently documented replication selection process can help characterize - and signal potential points of improvement for - a field's maturation.

To facilitate such a transparent reporting of considerations that led to a replication study, we aim to develop a list of criteria generally regarded as important, which could be used to systematically and transparently justify the selection of a particular replication target. Researchers could use this list to transparently *and* systematically report their replication target selection process, and in turn meta-scientists could use these reports to characterise a field's development. A great example for transparent selection of a replication target was recently published by Murphy and colleagues [24]. While this is a useful start to justifying resource allocation, we believe that we could go a step further by streamlining this process and offering authors structure and guidance in their selection process. Additionally, a list of considerations would offer a structured tool to funding agencies both when providing money for replication studies specifically and when looking to evaluate the usefulness of a proposal. In the remainder of this paper, we outline how we plan to go about developing this list.

(a) The present study

We argue that the involvement of the wider scientific community is crucial when designing a list of considerations to be used for transparent and systematic replication target selection. In this project, we aim to (1) describe the considerations generally regarded as important by psychological researchers and (2) construct a list of considerations to be consulted when selecting future replication studies in psychology. To ensure that our results reflect considerations of the selection process generally regarded important by the psychological community, we will employ

a consensus-based method.[1] More precisely, we will use a Delphi approach to expound the considerations and criteria researchers commonly deem important when selecting a replication target.

The Delphi process, which has the goal of developing consensus on a given topic or issue, is one of the most frequently used methods across multiple fields [25]. The Delphi process, as applied in this setting, is descriptive and can be considered an exploratory sequential mixed methods design. It is an iterative process, in which judgements from 'informed individuals' are collected in the form of questionnaire responses. The questionnaire collects both quantitative data in the form of importance ratings and qualitative data in form of suggestions and opinions on judgements. Over several rounds, consensus on several judgements or opinions emerges [26]. We have chosen this method for use in the current project, as it allows for including researchers from all over the globe, ensures anonymity of responses which allows participants to disagree more freely [25], and is most likely to yield results which reflect the opinions of the group as a whole, rather than capturing the views of a select few outspoken individuals.

We will implement a so-called 'reactive' Delphi method [26]; a modification of the original Delphi method. The reactive Delphi method involves participants responding to a previously constructed version of items, instead of generating a list of items themselves [26]. In the present study, a preliminary list of items was constructed by the organizing authors (M.-M.P, S.M.F, P.M.I, A.E.v.t.V., and D.v.R.) before registration of the project. The organizing authors combined elements from previous suggestions on how to justify replication target selection [e.g., 18,19] to create a preliminary list of considerations.

A disadvantage of this method is that the quality of the resulting consensus largely depends on the quality of the questionnaire design [i.e., the initial list of considerations; 25]. The authors acknowledge that they might have missed some crucial considerations when constructing the preliminary list of considerations. To overcome this,we will include an additional survey round with individuals who selected a replication target in the past. Participants will be asked to report how they selected replication targets in the past before judging the preliminary list of considerations. Additionally, participants will have the opportunity to suggest additional considerations not yet included. We will use the information from the survey to adapt our list of considerations. With this extra step we hope to ensure that the questionnaire sent out to the informed individuals contains all relevant elements.

Additionally, the survey enables insight into the specifics of the selection process and whether it differs depending on the researcher's motivation for conducting a replication and the type of replication. Different considerations might apply to replications that can be more readily termed close replications (e.g., more methodological) than to replications that are more conceptual (e.g., more theoretical). Mapping researcher considerations onto the different types of replications will bring the field one step closer to more explicitly matching outstanding questions for a specific phenomenon with the type of replication that most efficiently answers them (e.g., if an original result is expected to be a false positive, a close replication might be the best match). Although in reality there are many different forms a replication can take [see e.g., 27,28] the distinction between close (direct) and conceptual replication is most common and well-known by researchers, which is why for the current survey we examine the relationship with researcher motivations and these articulated ends of the continuum.

Lastly, the updated list of considerations will be send to a selected group of informed individuals, or 'experts', on replication target selection. Over the course of several rounds, participants will be asked to judge considerations based on their importance, and given the opportunity to suggest revisions. After each round, consensus will be evaluated based on pre-specified criteria and participants will receive a report summarizing the feedback from the previous round. For an overview of the proposed method, see Figure 1.

[1]We recognize that there may be much disagreement on a local level about what is important – we aim to characterize the opinions of researchers on average, to the extent that that is possible

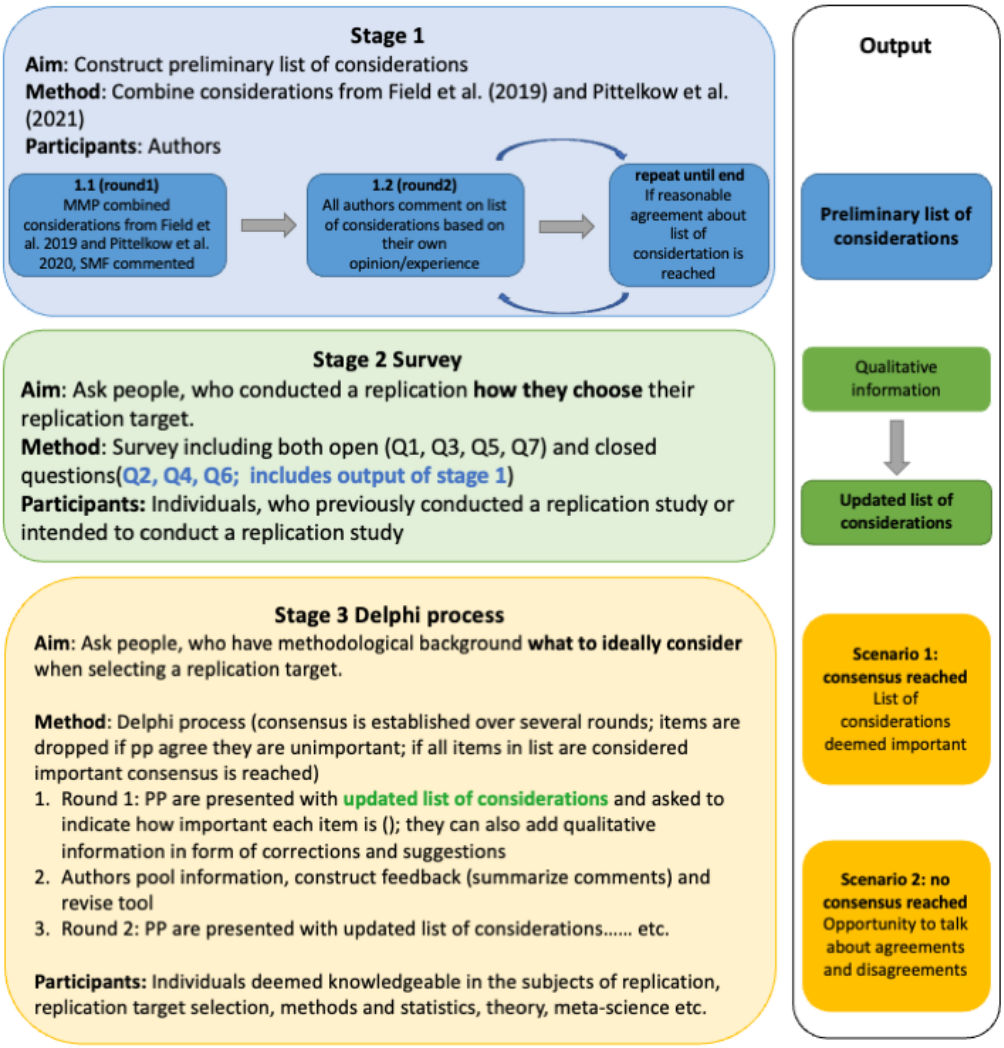

Figure 1. Flowchart illustrating the three stages planned for this project.

1. Methods

Ethics. Ethical approval for the proposed method was granted by the Ethical Committee Psychology (ECP) of the University of Groningen, the Netherlands on 04/02/2021.

(a) Researcher Description

M.-M.P. has previously published work on replication target selection in clinical psychology [19]. Her interest in the topic stems from a background in clinical psychology and the realization that sometimes "shaky" effects are translated into clinical practice. In her opinion, (1) treatments should be recommended only with sufficient evidence, also achieved by replications, and (2) which studies to replicate and how should be determined by evaluating a set of candidate studies. D.v.R. has published theoretical work on replications [18,19,29] and has conducted empirical replications [30,31]. S.M.F. has also published theoretical and empirical works concerning replication [18,30]. Frustration with (sometimes) inefficient use of resources and insufficiently justified reasoning behind conducting replications drives her interest in providing researchers

with the means to help systematize the replication target selection process, which can be difficult to navigate. P.M.I has previously authored theoretical work on replication target selection [16,20]. A.E.v.t.V has previously published about theoretical and practical aspects of conducting replications [32], has conducted large scale and Registered Replication projects [9,33–35], and is involved in theoretical and meta scientific work on replication target selection [16]. Her experience in analysing replications within psychology strengthens her belief that more explicit characterising of (the process of conducting) replications and their various functions can be a step towards making replication common in research lifecycles and towards theory building through conducting progressive types of replications.

(b) Stage 1

This stage was performed before registration. We created a preliminary list of factors to consider when deciding what to replicate from [18] and [19]. In a first step, author M.-M.P. extracted themes from these previous publications, and grouped them according to the four themes (1) uncertainty, (2) value/impact, (3) quality, (4) and cost/feasibility identified by [16]. Next, author S.M.F. commented on the list and author M.-M.P. adapted it accordingly. Lastly, the organizing authors (M.-M.P, S.M.F, P.M.I, A.E.v.t.V., and D.v.R.) provided feedback and the list was adapted over four rounds until all authors agreed on the final list of 16 considerations. Starting in round three, the authors agreed to not group considerations according to the four themes, as multiple themes applied for some considerations. For example, items grouped under quality, such as sample size, could also inform uncertainty. We will however, after the next stage, ensure that all initial and additional themes will be represented in the pool of items. This first list of considerations can be found in Table 1 and the process file is available on OSF.

Table 1. Preliminary List of Considerations Constructed in Stage 1 and the Corresponding Item Number for the Stage 2 Survey.

Nr.	Consideration	Corresponding Item *
1	Do you consider the current strength of evidence in favor for the claim to be weak (as for example quantified by a Bayes factor, a very wide CI, or a p-value close to the typical alpha level of 0.05 combined with a very large sample size).	Q7 item 12
2	Given the current state of investigation of this claim in the literature, how certain are you that the claim is true? Please motivate your answer.	Q5 item 5
3	Is the claim theoretically important? If yes, please elaborate.	Q5 item 4
4	Do you perceive this claim to have relevant implications, for instance in practice, policy, or clinical work? If yes, please elaborate.	Q5 item 3
5	Please describe the design of the original study.	Q7 item 2-5
6	Enter the sample size	Q7 item 1
7	Who was the sample (for example, what were inclusion and exclusion criteria)?	Q7 item 2
8	How was the main outcome measured?	Q7 item 19
9.1	Do you consider the outcome measure to be valid? Please motivate your answer.	Q7 item 9
9.2	Do you consider the outcome measure to be reliable? Please motivate your answer.	Q7 item 10
9.3	Do you consider the outcome measure to be biased? Please motivate your answer.	Q7 item 11
10	Do you consider the operationalization appropriate (i.e., are the methods fitted to answer the broader research question that was posed)?	Q7 item 20
11	Please describe the analysis plan and performed analysis.	Q7 item 13, 14, 17
12	Please enter the observed effect size	Q7 item 6,7
13	Given the sample characteristics, was the sample a good representation of the population? In other words, do the results generalize to the population of interest?	Q7 item 8
14	Is the interpretation of the current claim limited by potential confounds? If yes, please describe	Q7 item 21
15.1	Given the original study set-up, is replication readily feasible?	Q9 item 2
15.2	Can this study be replicated by generally-equipped labs, or are more specific experimental set-ups necessary (e.g., an eye-tracking machine, an fMRI-scanner, a sound-proof booth, etc.)?	Q9 item 1
16	How could a replication overcome the issues you raised above? Please also specify the type of replication you intent to run (i.e., close or conceptual).	

Note: *Item numbers refer to the presentation in the supplement

(c) Stage 2

(i) Participants of the Survey

We sampled psychological researchers who previously selected a replication target, identified as having either conducted or registered a replication study. We contacted individuals identified through a systematic review of the literature and online search.

We developed a search strategy via pilot searches documented in the supplementary material (i.e., *methods and additional analysis*). Similarly to previous studies [36,37], we identified potential participants by searching the following categories in *Web of Science* using the search string *TI = (replication OR replicated OR replicate)*: Psychology Biological, Psychology, Psychology Multidisciplinary, Psychology Applied, Psychology Clinical, Psychology Social, Psychology Educational, Psychology Experimental, Psychology Developmental, Behavioral Sciences, and Psychology Mathematical [2]. We refined time-span to the last five years. To overcome publication bias, we additionally searched the OSF registries using the term *replication OR replicated OR replicate*, again focusing on psychological studies registered in the past five years.

Contact information of corresponding authors from eligible articles was extracted. Articles and registrations were eligible if they concerned either a close replication or a conceptual replication in the field of psychology. We defined replications as projects concerning the same effect/hypothesis, independent, and dependent variables as specified previous work [27]. In judging eligibility, we mostly relied on the authors self-presentation. We excluded (1) student projects, as it is unclear whether the replication target was selected or assigned, (2) studies which were clearly not psychology, (3) hits that did not correspond to a research paper or registration, and (4) projects not identified as replications. We were lenient in our exclusion criteria as we expected some self-selection on the side of the participants. This means that we also contacted authors of work where we were unsure whether inclusion criteria were fully met. For eligible registrations, we searched for potential research output and extracted contact information from those records. If no research output was available, we noted (1) the author of the registration, or (2) the author of an associated OSF project (in that order).

The screening procedure is illustrated in Figure 2[3] and a full overview is provided on OSF. If the same corresponding author was identified multiple times, we (1) selected the project with clearly met eligibility criteria over one where we were unsure, and (2) selected the most recent project (i.e, the one for which the decision was most recent), as we assumed that participants would be best able to recall the selected process for most recent projects. In one case, we identified eight projects from one author all published in 2021. In this case, we select one project randomly.

Some of the participants were distant colleagues of the research team. However, the authors did not interact with participants as data was collected anonymously online. Nonetheless, the author names were disclosed during the survey, which might have impacted data collection.

(ii) Sample Size

The survey in stage 2 served as a pilot to inform the list of items provided for the first round of the Delphi process. Sample size determination for qualitative work is complex and depends on a variety of factors such as the scope and nature of the research, the quality of the data collected, and what resources are available. Here, we based sample size considerations on the available pool of potential participants. Typically, qualitative studies report between 20-30 participants. For the purpose of our project, we deemed it crucial for our sample to be large enough to be reflective of the consensus in the field. We identified a total of 682 potential participants and with a response rate of 10% we expected our sample to be twice as large as recommended..

[2]Some differences in label terms from [36,37] are due to Web of Science updates
[3]Please note that Figure 2 contains a correction as some duplicates were identified after receiving IPA

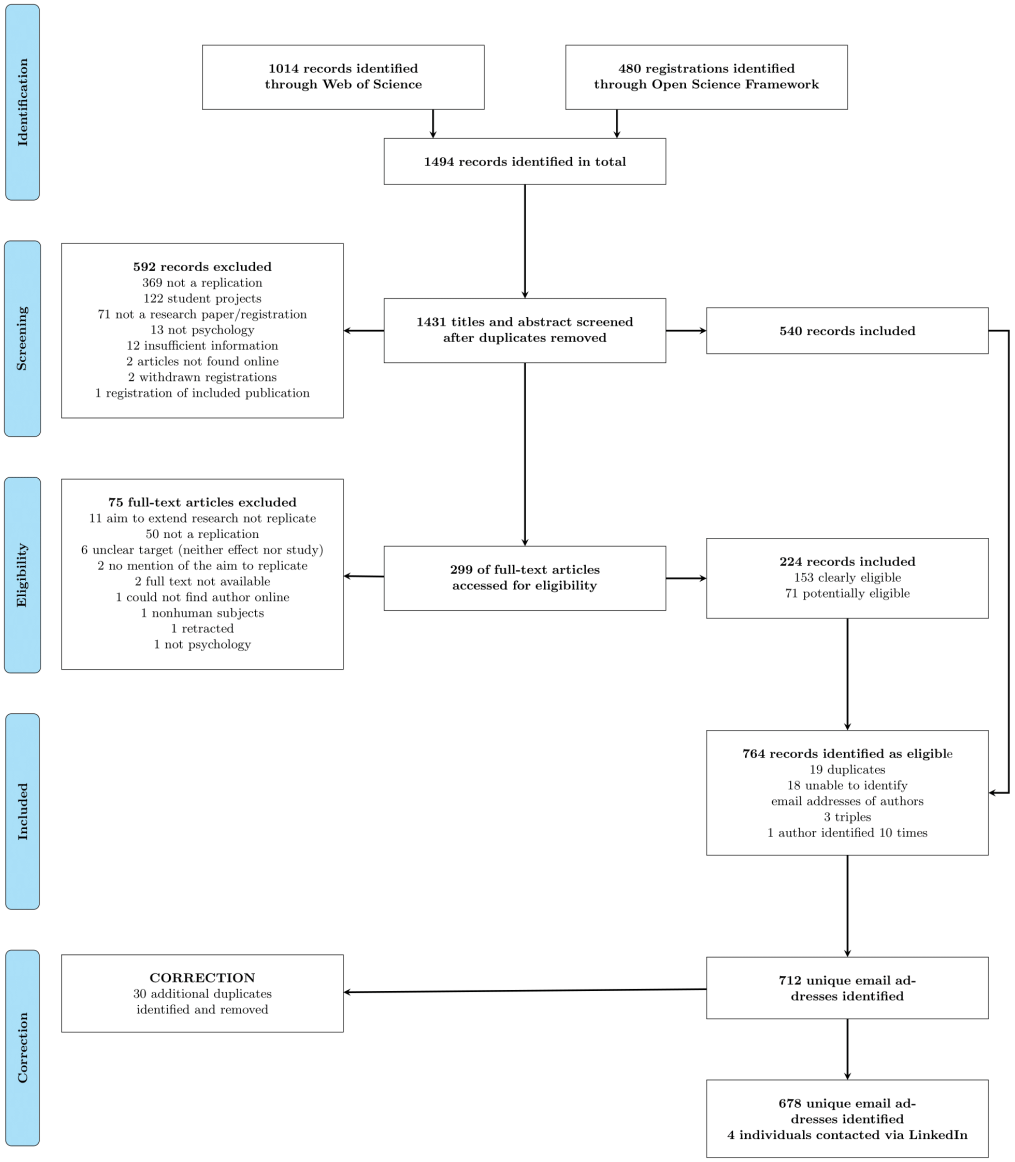

Figure 2. Flowchart illustrating the identification of potential participants.

(iii) Procedure

To gain insight into the replication target selection process and pilot the preliminary list of items, we constructed an online survey with eleven questions. The aim of the survey was to (1) pilot the considerations included in the preliminary list and those the author group was undecided about, and (2) to capture considerations not mentioned in the preliminary list. The former was achieved through closed questions rated on Likert scales, and the latter through open questions leaving room for suggestions and additional information. The survey questions are detailed in the supplementary material (*materials and additional data analysis*).

First, we asked researchers to identify the psychological field they work in (closed question). We adapted the sub-field choices from [38] and [37] and offered participants the choice between: Cognitive and Experimental Psychology, Clinical and Personality Psychology, Developmental

and Educational Psychology, Industrial and Organizational Psychology, Biological and Evolutionary Psychology, Neuropsychology and Physiological Psychology, Social Psychology, Quantitative and Mathematical Psychology, Human Factors, Unsure, and Other.

To gain insight into how replication targets are selected in practice, we asked our participants to illustrate what motivated them to replicate and how they came to pick the particular replication target they chose (open question Q2). Next, participants were asked to describe the type of replication they conducted (open question Q3) and self-identify as either close, conceptual, or other (closed question Q4). To probe our initial list of considerations, we asked participants to indicate to what extent they considered general study characteristics of the OS (closed question Q5), specific study characteristics of the OS (closed question Q7), and feasibility of a potential replication study (closed question Q9). For each of these three aspects, we presented a number of items. On a Likert scale ranging from 1 (*not important at all*) to 9 (*very important*) participants were asked to *"indicate to what extent* [they] *considered the following pieces of information"*. Items represented the initial list of considerations as well as aspects that the authors did not agree upon during stage 1, but which were considered *very* important by at least one author.[4] To avoid ordering effect, items were presented randomly to each participant, such that each participant their item list in a different order. After each closed question, participants had the opportunity to provide "*any other considerations you had with respect to general study characteristic*" (Q6) "*specific study characteristics*" (Q8) or "*feasibility*" (Q10). Lastly, Q11 provided the opportunity to give general comments and feedback on the survey. To counteract missing data, participants were prompted if they did not answer a question.

Candidate participants identified through the systematic review were contacted via email including a short description of the project and a link to the online survey. The contact email can also be found in the supplementary material (*methods and additional analysis*). We estimated the survey to take approximately 15-20 minutes. Data collection was open for a month and reminder emails were sent one and three weeks after the initial invite.

(iv) Data Analysis Plan

Open-ended items (i.e., Q2, Q3, Q6, Q8, Q10) were analyzed using thematic analysis. Thematic analysis is used to identify patterns (themes) within data [39]. During thematic analysis, the researcher plays an active role in identifying, selecting and reporting themes [39]. For the purpose of our project, we used thematic analysis as a realist method, reporting on experiences and judgements from our participants. In contrast to quantitative methods, qualitative data analysis is an inherently flexible and exploratory process, Braun and Clarke (2006) mention a number of questions one can consider before data collection, on which we reflected here:

- Themes were identified at the semantic level meaning that we focused on what was explicitly mentioned in the data without examining the underlying ideas and assumptions which shape the content. As such, we consider our analysis to be descriptive.
- We used an inductive, data-driven approach for identifying themes. To this end, we read and re-read the data for themes related to considerations for replication target selection. We were aware that our previous involvement with the topic might impact the themes identified and aimed to be reflexive during the coding process [40, c.f.]. Reflections and potential sources of bias were documented. Relevant text from these reflections is discussed in the manuscript, while details can be found on OSF.
- We were interested in extracting the most frequently mentioned themes (i.e., considerations). Prevalence was counted across and not within individuals. In other words, we counted how many individuals mentioned a certain theme, and not how often the theme was mentioned overall. When registering this report we consciously refrained

[4]Item 16 of the initial list of considerations was the only item not included, because it does not feature a unique consideration for choosing a study to replicate.

from quantifying the proportion of participants that need to mention a theme for it to be considered frequent, so that later we were able to judge which themes are the most crucial ones, and in which proportion based on the data. In the result section we report the number of instances themes were mentioned across individuals.

- We were interested in comparing themes between different types of replications. Thus, we contrasted codes identified in responses to Q2, Q6, Q8,and Q10 between different types of replications identified in Q3 and Q4.

Braun and Clarke (2006) suggest that thematic analysis consist of six phases: (1) familiarization with the data, (2) initial code generating, (3) theme searching, (4) reviewing themes, (5) defining and naming themes, and (6) producing the report. These phases are not to be performed one after the other; instead the data analysis process is recursive, with the researcher moving back and forth between these phases. Our approach was similar to these broad guidelines. It involves the following preregistered steps:

First, authors M.-M.P. and S.M.F. split the data in two, and worked independently on developing a set of likely codes based on themes identified in the data at this stage. Our approach in this step was consistent with the practice of open coding, that is, we selected chunks of relevant text and associated them with a short phrase or keyword generated from the text itself. Second, authors M.-M.P. and S.M.F. collaborated with one another to determine which codes to include in a codebook. This codebook contained information for each code including a thorough definition of the code itself in the abstract, text snippets as concrete examples, and descriptions of inclusions and exclusions (i.e., concrete cases where a given code might not apply). The codebook is openly available on OSF.

Once the codebook was established, authors M.-M.P. and S.M.F. each went through the qualitative text in its entirety, and coded it according to the codebook. Our unit of analysis was a sentence. Once each person coded the dataset, interrater reliability (IRR) was calculated.

According to Miles and Huberman [41], IRR can be calculated as the total number of agreements (between authors M.-M.P. and S.M.F.) divided by that same numerator, plus the number of disagreements between authors M.-M.P. and S.M.F.. Miles and Huberman suggest that an agreement rate between coders of 80% is sufficient, and we used this same threshold. We had planned to consult a third author (i.e., AvtV), if we had not reached the anticipated IRR. That is, as Syed and Nelson put it, "one individual's analysis of qualitative data should generally lend itself to be re-captured by another individual who is reasonably familiar with the research question and procedure" (p. 376) [42]. Although replicability is arguably difficult to apply in the context of qualitative research, consistency between coders in this case can certainly be validly applied. IRR performed as a measure of our consistency. Final steps in this process revolved around reviewing, defining and naming themes, as Braun and Clarke suggest.

Closed-ended items were evaluated using the median rating and interquartile range (IQR), a measure of dispersion around the median capturing the middle 50% of observations [43]. Old items with a median rating of 3 and an IQR of 2 or lower were excluded from the list, and new items with a median rating of 7 and an IQR of 2 or lower were included. To explore whether the considerations differed between field of expertise and type of replication, we stratified the sample and compared subgroups.

(d) Stage 3

(i) Participants of the Consensus Process

Panel members were identified using snowball sampling, a type of convenience sampling. Snowball sampling is one of the most frequently employed methods of sampling for qualitative research [44], and especially useful if participants need to meet specific criteria or have certain expertise [45]. First, the research team identifies a number of potential candidates. Next, the identified people are contacted and asked to participate and/or identify others who they see fit

to participate in the study. By asking potential participants to consider who else has the expertise needed for the study, snowball sampling taps into social knowledge networks [44], which we considered beneficial to our project as we were interested in shared, communal knowledge regarding replication target selection.

Snowball sampling was implemented as follows: Prior to registration, we constructed an initial list of 29 potential participants, who we deemed knowledgeable in the subjects of replication, replication target selection, methods and statistics, theory, or meta-science. The list can be found in the supplementary material (*methods and additional analysis*. To not only identify "replication experts", but also content researchers, we offered researchers who participated in the stage 2 survey the option to sign up for the Delphi procedure.

Next, we contacted these potential candidates via email, asking them whether they were willing to participate and/or to forward the invitation to someone they might find eligible, and/or to nominate another person by replying to the email. We are aware that this method does not ensure that every potential participant has an equal chance of being selected. To avoid the sample being heavily biased, we attempted to balance participant selection regarding gender, career level, and country of residence. We planned to make a Twitter call to reach out to members of underrepresented demographic category, relying on 'word of mouth' in the scientific community on Twitter if necessary [5]

Eligible participants received an online survey, asking them to indicate their agreement with the previously constructed list of considerations on a Likert scale from 1 (*not important at all*) to 9 (*very important*). We also offered the option for free text responses on the phrasing of the considerations and whether important considerations were missing. Quality of consensus is highly dependent on participant motivation. To ensure that our participants were sufficiently motivated, we offered co-authorship in exchange for participation. Authorship was voluntary and not a prerequisite for participation[6]. If Delphi experts decided to identify as authors they were considered *investigators* according to the CRediT taxonomy (see Authors' Contributions).

We anticipated the sample to consist (mostly) of researchers who are distant colleagues or perhaps one-time collaborators with some of the author team. Our contact with them in the context of the study was distant.

(ii) Sample Size

Some authors suggest a sample size around 20 members to produce stable results [46,47], while others argue that smaller panels of 6-11 panelists suffice [25]. However, individual responses are very influential in small panels producing potentially unstable results [46]. As the Delphi process is time-intensive, panel attrition is likely. Typically, the overall response rate for Delphi procedures is 80% [48]. Thus, we aimed to recruit a minimum of 30 participants for our study over a maximum period of three months.

We planned that if after 1 month, our sampling procedure resulted in more than 30 participants, we would proceed to the Delphi process, provided that the sample was balanced with regard to gender, career level, and country of residence. Additionally, we planned to stratify participants by their research field similar to [37]. Otherwise, we decided to reject and select participants to create a balanced sample. In the latter case, we planned to report justifications for participant selection. We further planned that if, after three months, our sampling procedure resulted in fewer than 30 participants, we would proceed with the Delphi process but highlight that results might be unstable and recommend replication to establish stability of the considerations. Please note that the sample size determination was empirically informed as no clear guidelines for 'optimal' panel size for Delphi procedures exist.

[5]We acknowledge that such an approach may introduce selection imbalances of its own, however we argue that it is still likely to assist in reaching a wider range of participants.
[6]One participant opted to not be listed as a co-author

(iii) Procedure

The goal of a Delphi process is to establish consensus over several, iterative rounds. During each round, participants were asked to judge the importance of a number of items (i.e., considerations) and provide feedback. In between the rounds, participants received structured feedback reports summarizing results from the previous round both quantitatively and qualitatively.

For the first round, participants received the list of considerations, updated by the results of stage 2. For each subsequent round, participants received a revised list of items, for which consensus had not yet been reached. Items were revised according to qualitative feedback from the participants. To define what constitutes consensus and avoid the Delphi process going on indefinitely, stopping rules were implemented. In line with [38] the following pre-specified stopping rules applied: (1) the Delphi process was defined to be "*concluded with unsuccessful recruitment*" if three months after contacting potential panel members, there were fewer than 6 participants; (2) the Delphi process was defined to be "*concluded with consensus*" if consensus was reached about the considerations generally regarded as important when selecting a replication target. Consensus was defined as an IQR of 2 or less. Once consensus was achieved for all items, no new round would be initiated; (3) the Delphi process was defined to be "*concluded with incomplete consensus*" if consensus was not reached for all items (i.e., IQR > 2) after the fourth round. No new round would be initiated after this stopping rule was triggered. We planned to report the last version of list of considerations and highlight disagreements.

(iv) Data Analysis Plan

Data analysis was performed after each Delphi round. Quantitative items were analyzed using medians and IQR and the distribution of ratings were visualized using histograms. Items with a median rating of 6 or more and IQR of 2 or less were included in the final list of considerations. Items with a median rating lower than 6 and an IQR of 2 or less were excluded. Qualitative responses were summarized by M.-M.P. and discussed by the author group. We counted how many individuals mentioned a certain concern or suggestion. The list items were revised based on frequently mentioned suggestions. When registering this project, we consciously refrained from defining *frequently* a priori to allow us to flexibly respond to concerns and suggestions later on. We anticipated no incomplete data reports as we forced participants to answer every item. If participants had no suggestions, they were instructed to answer open questions with "*none*". If due to attrition, participants did not join subsequent Delphi rounds, we proceeded with the remaining experts.

After each round of data analysis, M.-M.P. constructed a structured feedback report for the participants. Items for which consensus was reached were not included in the summary report to the participants. In the feedback report we: (1) replied to frequently raised general concerns if there were any, and (2) presented items for which no consensus was reached. For each item we presented the histogram of responses, highlighted revisions if necessary, and addressed item-specific concerns. Summary reports and the invitation for the next round were sent to participants who responded to the previous round.

(e) Reporting of results

During stage 2, we produced: quantitative data (i.e., importance ratings), qualitative data (i.e, participants responses and corresponding codes), documents containing reflections and potential sources of bias from coding authors, and an updated list of considerations. Quantitative data was summarized using median ratings and IQR and is presented in tabular form. We report identified codes and associated frequencies. Reflections of the coding authors and the updated list of considerations are available at OSF. Reflections and reasoning behind what qualified as a theme are discussed in the manuscript, leading to intermediate conclusions about how psychological researchers select replication targets. During stage 3, we produced quantitative (i.e., importance ratings), and qualitative data (elaborations from participants), feedback reports for each round,

and a selection of items, which participants agreed upon (i.e., the final checklist), and potentially items that no consensus was reached for. Median ratings and IQR for each item across the rounds are presented in a table. We report our definitive checklist, highlighting in particular the items that reached consensus, but also those that did not. Feedback reports were uploaded to OSF, summarizing also the qualitative input from the Delphi process. These results allowed us to discuss and suggest relevant considerations for future researchers, discuss implications for psychological science, and potentially other social sciences and signal potential direction for future research.

2. Results

(a) Protocol and Data

All supplementary material including the pre-registered manuscript, which received in principle acceptance, data and analysis files can be found on OSF.

(b) Stage 2

(i) Deviations from preregistered plan

While we followed the pre-registered plan as closely as possible, a few deviations were deemed necessary.

Stage 2. First, during data collection 30 additional duplicate emails were identified and removed according to the pre-registered protocol. If we had identified two email-addresses for one person, we used both to increase the likelihood of a response. Second, despite repeated prompts for participants to answer all items, some data was missing. Some participants indicated why they were unable to answer specific items, thus providing us with qualitative information about the mechanism of missingness. We therefore considered responses with missing data on some, but not all items, as complete and included it in the quantitative analysis with all data available[7]. Third, we had planned that authors M.-M.P. and S.M.F. would collaborate first with one another, then with the other authors, to determine which codes to include in a codebook. However, the code-book was established by M.-M.P. and S.M.F. without the input of the co-authors. Codes overlapped substantially and disagreements were easily resolved. Lastly, while we meant to exclude all student projects when identifying potential participants, 16 participants indicated that they conducted their replication as student projects. Their responses were included in the analysis as we were committed to use all available data and the respondents were able to describe their decision-making process.

(ii) Participants

A total of 682 participants were contacted. Of these, 678 individuals were contacted via email on 04.10.2021 using the Google extention GMass. Four additional individuals were contacted by M.-M.P. via LinkedIn on 11.10.2021. Details about the reminders are described in the supplement. Data collection was closed four weeks after it had started (i.e., on 01.11.2021).

A total of 185 (27%) responses were recorded. Of these, 64 responses were incomplete, leaving a total of 121 (18%) responses.[8] Demographic information of the 121 responders is presented in Table 2.

(iii) Quantitative analysis

We calculated the median and IQR for all quantitative items. Results are presented in Table 3 and visualized in Figure 3. None of the items reached our pre-specified decision criterion of a median

[7] Incomplete responses (i.e., when respondents stopped after a number of items) were excluded from the analysis.
[8] The algorithm indicated 66 incomplete responses but two were marked incorrectly.

Table 2. Number of participants by their field of interest and type of replication the participant has conducted.

	Total	Direct/close replication	Conceptual replication	Other
N	121	94	17	9
Psychology field (% per column)				
Cognitive and Experimental	39 (32.2%)	30 (31.9%)	5 (29.4%)	4 (44.4%)
Social	29 (24.0%)	23 (24.5%)	4 (23.5%)	2(22.2%)
Clinical and Personality	12 (9.9%)	9 (9.6%)	3 (17.6%)	
Developmental and Educational	7 (5.8%)	6 (6.4%)	1 (5.9%)	
Industrial and Organizational	5 (4.1%)	4 (4.3%)		
Biological and Evolutionary	4 (3.3%)	3 (3.2%)	1 (5.9%)	
Quantitative and Mathematical	4 (3.3%)	4 (4.3%)		
Human Factors	2 (1.7%)	1 (1.1%)		1 (11.1%)
Neuropsychology and Physiological	1 (0.8%)	1 (1.1%)		
Other[a]	11 (9.0%)	10 (10.6%)		1 (11.1%)
Unsure[b]	(5.8%)	3 (3.2%)	3 (17.6%)	1 (11.1%)

Note: One person did not indicate what type of replication they conducted and thus excluded from the stratified counts.
[a]conservation/environmental psychology, differential psychology, experimental analysis of behavior, human-computer interaction, legal psychology, metascience, parapsychology, psycholinguistic, social and evolutionary psychology, and sociology
[b] behavior genetics, communication and media psychology, economic psychology, media psychology, neuroimaging, and sport and exercise psychology

rating no larger than 3 with an IQR no larger than 2 and none of the new items reached our pre-specified decision criterion a median rating no smaller than 7 with an IQR no larger than 2. Consequently, we did not change the preliminary list of considerations based on the quantitative analysis.

A second aim of our survey was to examine potential differences in considerations based on the field of expertise and type of replication. To this end, we split the data into different strata and compared the medians and spread of the data (IQR, min and max) for each stratum. The stratified analysis is detailed in the supplementary material (methods and additional material). No meaningful differences were observed between sub-fields. Ratings differed slightly between the different types of replication. For example, participants that classified the replication they conducted as *direct* or *close* rated generalizability (*Mdn* = 4), in- and exclusion criteria (*Mdn* = 3), and random assignment (*Mdn* = 3) lower than participants that classified the replication they conducted as conceptual (*Mdn* = 7, *Mdn* = 6, and *Mdn* = 6 respectively). This is most likely explained by the different aims underlying close and conceptual replication. That is, while close replications aim to verify previous findings, conceptual replications aim to generalize findings beyond, for instance, the original study's context or sample. However, participants who conducted a close/direct replication rated statistical error as unimportant (*Mdn* = 3), which is in contrast to the assumptions that the primary aim of close replications is to verify[9]. Nonetheless, differences between subfields and type of replication were not substantial enough to warrant specific versions of the list of considerations for each.

(iv) Qualitative analysis

First, we split the data in half using 60 randomly generated numbers between 1 and 121. S.M.F. and M.-M.P. independently established codebooks based on 60 and 61 responses respectively. S.M.F. identified 56 codes, and M.-M.P. identified 67. In two consecutive meetings, S.M.F. and M.-M.P. reviewed and compared their codes and collaboratively established a codebook including 73 codes. Lastly, both S.M.F. and M.-M.P. independently re-coded the complete data set using the established codebook.

[9]We received qualitative feedback suggesting that some participants might have misunderstood this question. They meant to indicate that flawed studies should not be replicated (answer: no), where the question aimed to assess whether a study being flawed is a relevant factor for deciding to replicate (which for the above would mean, answer: yes). This limits interpretability of this particular item

Table 3. Survey questions with descriptive statistics used for quantitative analysis

Question	*N*	Median	IQR
Please indicate to what extent you considered the following pieces of information when scrutinizing the potential replication target:			
- Whether the finding has been investigated sufficiently or not.	119	8	3
- Whether the citation count of the study was high or low.	119	4	5
- Whether the study has relevant implications, for instance in practice, policy, or clinical work, or not.	120	7	3
- Whether the finding has a strong connection with theory or not.	120	7	3
- Whether the finding was unexpected (e.g., "counter-intuitive", "surprising"), or in line with what can be expected.	119	6	4
Please indicate how important the following specific characteristics of the original study were for you when choosing your replication target:			
- The total sample size.	115	6	4
- Handling of inclusion and exclusion criteria.	115	4	5
- Blinding procedures (e.g., blinding of participants, experimenters, analyzers).	117	2	5
- Sampling procedures (e.g., stratified random sampling, snowball sampling, convenience sampling etc.).	115	4	4.5
- How participants were assigned to conditions (e.g., randomly, single/double blind, etc.).	116	3	5
- Statistical power to detect the effect sizes of interest.	116	6	4.25
- The size of the effect size.	119	6	4
- Generalizability of the sample.	116	5	4
Validity of the outcome measures.	116	6	3
- Reliability of the outcome measures.	114	6	4
- Potential bias of the outcome measures.	115	5	5
- The strength of evidence (measured by reported p-value, confidence interval, Bayes Factor, etc.).	117	7	2
- Missing data handling.	114	3	4
- Whether the finding was based on within-subject measurements or between-subject measurements.	116	3	4
- Open access to underlying empirical data that were analyzed.	117	3	4
- Whether the study has been preregistered.	118	2	4
- Whether the finding was predicted a priori or discovered during data exploration.	114	5	5
- Whether there are statistical errors in the results reported (e.g., the degrees of freedom do not correspond to the other reported statistics, the total sample size does not equal the sum of the group sample sizes, etc.).	114	4	5
- How the main outcome was measured.	116	6	4
- Whether the operationalizations were appropriate (i.e., the methods were fit to answer the broader research question that was posed).	115	5	4
- Whether interpretation of the results was limited by potential confounds or not.	114	6	4
Please indicate how important the following pieces of information were for you when judging the feasibility of your replication study:			
- Whether the study could be replicated by a lab without specialised equipment (e.g., an eye-tracker, a sound-proof lab, an MRI-scanner).	119	7	5
- Whether the study concerned a hard-to-collect sample.	118	7	5

IRR was calculated as the number of agreements divided by the sum of the number of agreements and disagreements. Agreement was defined as both coders assigning the same code(s) to the same text or assigning the same code to different, but related, text. Disagreement was defined as both coders assigning different code(s) to the same text.[10] The first author noted cases of agreement and disagreement by going through the data case by case and (1) noting clear agreements (same code(s), same text), (2) noting unclear agreements (same code(s), different text), (3) noting clear disagreements (same text, different code(s)), and (4) noting codes only assigned by one coder. A detailed account of this procedure is provided in the supplementary material (*method and additional analysis*). In total, 343 agreements (1), 77 disagreements (2, 3), and 329 quotes identified by only one coder (4) were counted. This resulted in an IRR of 0.82. [11].

The large number of quotes assigned by only one coder might be explained by (1) differences in coding styles (M.-M.P. assigned many more codes than S.M.F. in general), (2) differences in involvement in developing the codebook (M.-M.P. was more involved than S.M.F.), or the coder being more familiar with their own codes as opposed to the one established by the other. The assignment of codes to text involves the interpretation of those texts by the coder; the observed discrepancies are neither surprising nor cause for concerns about validity. To be sure, as Braun

[10] If coders assigned multiple different codes to the same text, this was counted as one disagreement
[11] If disagreements including multiple codes were counted as multiple, IRR dropped to 0.77

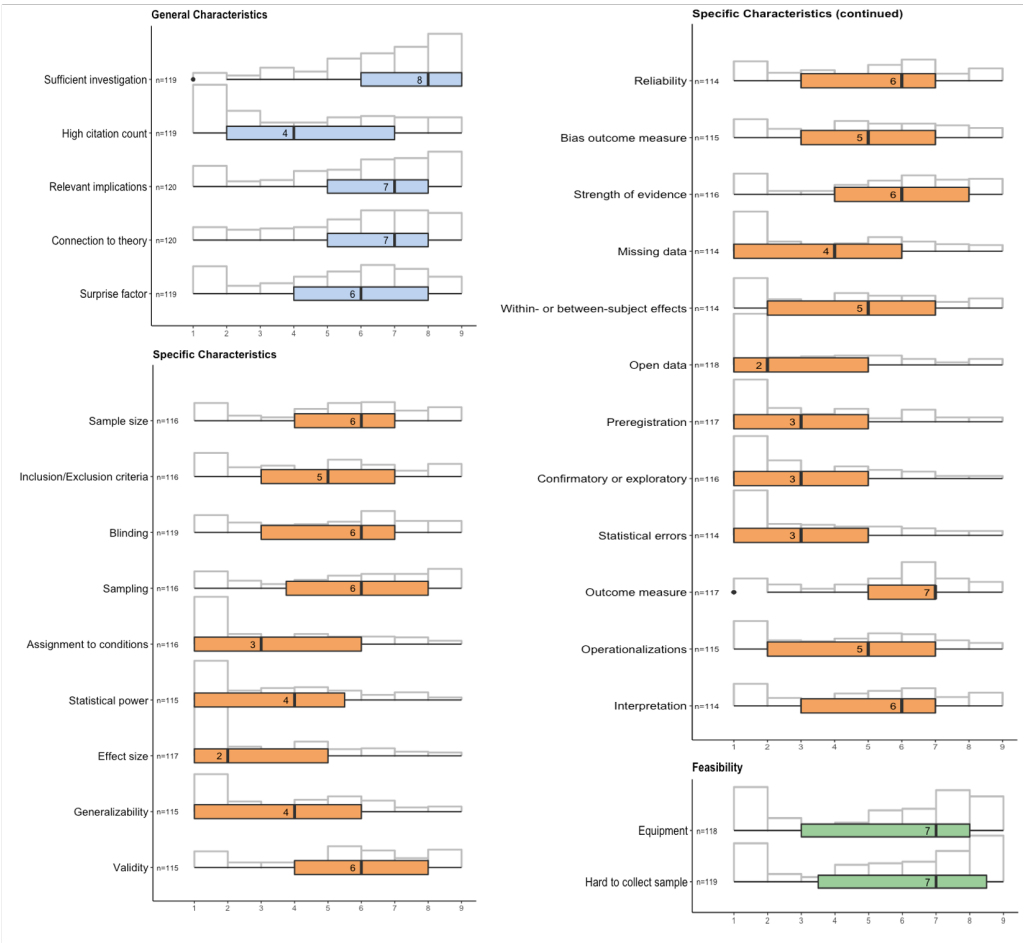

Figure 3. Quantitative Results.

and Clarke [49] emphasize, when multiple coders are part of a thematic analysis, the goal is to "collaboratively gain richer or more nuanced insights, *not* to reach agreement about every code." (p. 55, emphasis in original)

The coders identified two key themes. The first theme, **decision-making process**, describes the process underlying a participant's decision to replicate. In our interpretation, this theme is concerned with *how* participants decided to replicate, and encompasses the aids and obstacles they encountered during the process. The second theme, **motivation**, is concerned with *why* participants chose to replicate a study in general or *why* they choose their specific targets. Themes are not as distinct as we might present them in this text. Motivating factors interact with the decision-making process and vice versa. Below, we describe the themes and their specific sub-themes and relate them to each other. Participants' quotes are presented to illustrate themes and **themes** and important ***sub-themes*** are presented in bold and italics.

The decision-making process. To understand *how* our participants decided to replicate an original study, we coded their ***process***. We distinguished whether they decided to replicate based on a particular study or whether they decided to replicate before searching for a replication target. However, only 31 participants explicitly described their decision-making process. Moreover, this code was more frequently assigned by M.-M.P. than S.M.F.. Interpretation of these results is therefore limited. The mismatch in assignment frequency may reflect M.-M.P.s specific interest in

the process of replication target selection. Participants seemed to more frequently (n= 20) decide to replicate *after* reading or conducting a specific study than they decided to replicate before searching for potential replication targets (n=11).

Institutional influences shaped the decision-making process for some researchers (n=13). Four participants reported being invited to partake in larger replication projects, two of which did not describe their decision-process, presumably as others had made the decisions for them. Other respondents mentioned deciding to replicate for publication purposes. Three explicitly reported changes in journal policies regarding the publication of replications as motivators for them to conduct a replication. Specifically, they replicated original studies previously published in outlets that subsequently incentivized replications. For example, one participant reported that the OS they were interested in replicating was published in a journal that "*had recently adopted policy to publish pre-registered replication attempts for their own articles*" (Case 29) as one factor influencing their choice to conduct a replication.

Feasibility played an important role in the decision-making process of our participants (n=76), or as one participant put it "*feasibility was a key issue*" (Case 115). Feasibility refers to the ease of adapting (if needed), and running the replication study based on clarity and complexity of the OS, as well as the available resources. Feasibility was considered at different points during the decision-making process. For some, feasibility considerations preceded others, meaning that they only considered original studies which they could run based on their available resources. For example, one participant mentioned that "*[they] first considered whether [they] had the skills and resources to run the study*" (Case 104). For others, feasibility followed other considerations "*After that I selected studies with procedures for which direct replication would be feasible*" (Case 73). In this way, feasibility was used as a criterion to identify possible replication targets from a pre-selected pool of studies.

To determine the ease of conducting a replication, participants considered whether the method was sufficiently clearly described, and whether implementation of the OS was possible. For some, "*the study needed to have sufficiently detailed description[s] of [the] procedure, instruments and data analysis plan*" (Case 14). This sometimes coincided with participants mentioning the complexity of the original study's method, or more specifically, the ease with which the OS could be replicated. Participants seemed to look for "*methods [that] were clearly described and easy to implement*" (Case 82). However, not only studies with sufficient detail were replicated. For example, one participant reported that they "*did not realize how many information about the methods and materials was lacking in the paper*" (Case 79) until they conducted their direct replication. For some insufficiently provided information were a reason to refrain from direct replication but do "*partial replications because the Method section in the original study wasn't clear enough on some specifics*" (Case 94).

Participants further considered the ease with which they could adapt the OS. A few participants specifically mentioned that their replication target was "*easily extendable to additional condition [and], so it was a good fit*" (Case 81). One specific adaptation considered was whether the OS "*could be translated into other languages or cultural contexts*" (Case 93). While one might expect this consideration to be more prominent for conceptual replications, it was mentioned in relation to both direct and conceptual replication types.

Related to ease, participants frequently (n=18) mentioned the mode of data collection. Some participants specified the type of data they wanted to collect (e.g., questionnaire or performance data), but participants most frequently mentioned considering whether data was collected on location (e.g., a school or a laboratory) or online, and whether they could adapt the data collection. The need for online data collection was mentioned either as part of the OS methodology "*we only considered studies that were run online*" (Case 28), or as a possible adaptation "*adapting the method from an in-person context to an online/computerized setting*" (Case 82). Online data collection might have been a specifically relevant consideration in the context of the COVID-19 pandemic, which prevented many researchers from collecting data on site. For example, one participant specifically mentioned that they "*ensured it [the replication study] could be run online, in covid*" (Case 119).

Lastly, resources played a large role in considering the feasibility of potential replication targets. Participants considered the degree of overlap between available resources (e.g., time, money, available data, equipment, skills and expertise, and potential collaborators) and the resources required to replicate a specific OS. Participants frequently mentioned time constrains, meaning that "*[the replication study] had to be something that I could actually conduct given time and resources*" (Case 18). Time constraints were often mentioned in relation to financial constraints. Participants either discussed the need to find studies, which could be replicated at "*low costs*" (Case 75), the need to "*secure enough funding to make it [the replication study] happen*"(Case 22), or having "*the funding to support the replication*" (Case 107). Having access to the data, materials, and/or a participant pool, and potential collaborators who would be able to carry out the replication study eased the decision to replicate a specific target study. Lastly, some participants specifically mentioned considering whether they had the skills and expertise to replicate a specific target study. As one participant put it "*it was important that I had the expertise to perform the replication*" (Case 107).

Other infrequently mentioned aspects were the ease of getting ethical approval (n=2), participant burden (n=3), and whether the study ought to be multi-sited (n=3).

Naturally, the aspects of feasibility considerations were not mutually exclusive but overlapped within individual participants. For example, available resources would ease adaptation and adjustment of potential replication targets. As one participant described: "*I already had the software for the task, so it was pretty easy to adjust it for the new study*"(Case 5).

Motivation Participants' selection of replication targets was motivated by the replicating authors (RAs) **interest** in the original effect, **impact** of the original finding (perceived by the RAs, or objectively demonstrated, e.g., by citations or journal impact factor), **doubt** in the specific effect, specific **methodological aspects** of the OS, or was related to the **author of the OS**. In our interpretation, most participants were motivated by learning from the replication study. For example, five respondents conducted replication studies to gain familiarity either with the research process (e.g., Case 14), or the specific field of research one has not yet encountered (e.g., Case 18).

However, replications were not only conducted for personal benefit, but also for altruistic reasons. Ten respondents reported perceiving replications as good scientific practice and thus being committed to running them to "foster cumulative science" (Case 3) or "establish scientific credibility" (Case 55). Others (n=16) conducted replications for educational purposes either as seminar classes, theses, or joined research projects.

Interest motivated the majority (n= 83) of our participants to conduct a replication study. Many (n=32) specifically mentioned that (aspects of) the OS interested them and motivated their decision to replicate. Participants called it "*interest in the topic*" (Case 3) or simply stated "*the study we chose was interesting*" (Case 58), sometimes also labeling it as "*curiosity*" (Case 10). Three participants said they were interested in participating in the scientific discussion rather than aspects of the OS per se, and used involvement with a replication study to do so.

Participants mentioned several areas of interest, the most frequent (n=34) being the motivation to verify the literature body. Many (n=13) participants were planning to conduct their own experiments in the line of the OS, but wanted to verify the validity or reliability of the effect they aimed to extend first. Three other respondents were specifically interested in verifying the paradigm used in the OS as they were planning to use it in their own research. However, verification of the literature body was not always self-serving. Some (n=5) specifically mentioned the motivation to verify the literature body to foster knowledge or explore robustness of the effect. Respondents mentioning the motivation to verify the literature most frequently (n=30) conducted close replications. This is in line with our assumption that the function of close or direct replications is to verify existing research.

Related to, and overlapping with, the motivation to verify the literature body, many participants (n=19) reported an interest in self-replication. This meant that participants repeated

their own studies either because it was standard practice to them – "*Typically, we (our lab) provide replication studies *within* the original papers*" (Case 11) – or to verify their own findings. Verification could be motivated by methodological shortcomings. For example, one respondent noticed that "*the results was on shaky ground for some methodological shortcomings*" (Case 62). Most frequently, however, our respondents wanted to ensure that their findings were robust, valid, and stable.

If not their own studies, respondents were frequently (n=17) interested in replicating OSs that were relevant to their own line of research or that they were familiar with. One participant explained that replicating studies familiar to the researcher was attractive because it was relatively easy:"*had conducted a previous study with similar methodology and knew that [they] could easily do another, similar study*" (Case 118). However, mostly respondents opted to replicate studies that were "*influential to [their] ongoing research program*" (Case 82). Within one's line of research, interest was also sparked by novel methods, tools or measures. Sometimes, novelty coincided with "*striking*" (Case 119) findings. Other times, the OS "*broke very new ground*" (Case 96). As one participant put it "*we felt that something that novel and unexpected [. . .] should be replicated*" (Case 96).

Likewise, the context of the OS interested some respondents (n = 14). Participants were interested in context-dependency of the original effect or how changes in cultural and societal context might have impacted the original findings. For example, one participant "*was finding different results in another context and wanted to understand the phenomenon better*" (Case 34) and thus explored the context-dependency of the OS. Another respondent postulated that "*the results might be different in a sport context*" (Case 39). Similarly, some (n=5) respondents specifically mentioned interest in exploring the boundary conditions of the original effect.

Impact of the OS was mentioned by 61 respondents. Our participants replicated studies they judged to be generally important or "*seminal*" (n=24, e.g., Case 27, 87), to the field. For example, one participant explained that the replication target "*was a study that had had a considerable impact on our field*" (Case 61). Impact was sometimes defined as "*a lot of people talking about it*" (Case 81) or "*a lot of labs doing conceptual replications*" (Case 61), or a study pioneering a method not commonly used in the field of research. Overall, it appeared that our respondents were motivated to replicate cornerstone research, which was perceived as most valuable if the replication had impact regardless of the outcome.

Additional qualifiers of impact were citation count (n = 20) and the journal that the OS was published in (n=10). As one participant put it "*we choose to replicate [the OS] because: [. . .] it is an influential finding, as the original article is a well cited paper, published in a high impact journal*" (Case 85). Another respondent identified the OS as part of the scientific discourse, and therefore important to replicate, as it "*was published in a high ranking journal and [. . .] cited multiple times*" (Case 21). It appears that citation count and impact factor were used by many participants to judge the impact of an OS.

Studies were also identified as impactful by participants if the conclusions had theoretical relevance (n = 19). Replication was believed to "*provide insight into the credibility of [. . .] theory*" (Case 93) or enable participants to "*weigh in on a larger theoretical debate*" (Case 111). There was some discrepancy as to the role that theory played in the decision-making process. While theory could be regarded "*as unimportant, because presumably the theory that underlies replication targets is weak to begin with*" (Case 10), theory was also specifically mentioned to be "*powerful and [. . .] well specified/falsifiable*" (Case 7). It appears that there is no consensus as to whether studies with weak or strong theory ought to be replicated.

Eleven participants also considered the impact of the replication study instead of the OS. Respondents were motivated to replicate studies "*for which in the past no direct evidence was available*" (Case 4) or which were judged by them to be "*understudied topic[s]*" (Case 49). Respondents appeared to assume that replications could serve an important role if the evidence regarding the original finding was limited. However, one participant cautioned that "*a study may not be worth replicating simply because the phenomenon under investigation is understudied – there may be a reason why few studies have been conducted on a particular topic (e.g., little to no clinical or theoretical merit)*" (Case 106).

Impact outside of the academic discourse was also considered by nine respondents. Specifically, the impact of the original finding on society or policy and the public interest in the original finding. Though one mentioned that they did not care about policy implications (Case 109), the other eight were motivated by the practical importance of their replication study.

Doubt motivated 62 of our respondents to replicate a study. Doubt means that the RAs believed that they had reason(s) to be sceptical regarding the 'truthfulness' of the original finding. This was mostly (n= 22) due to potential flaws of the OS. Some respondents suspected the original finding to be "*due to design error or confound*" (Case 5) or "*the original study [to have] a series of methodological and statistical flaws that called the results of the original study into question*" (Case 13). As one respondent put it "*the [original] result was on shaky ground for some methodological shortcomings*" (Case 62), thus motivating replication to overcome said shortcomings. Interestingly, potential flaws were mentioned for both close and conceptual replications, though it stands to reason that in either case participants modified the original methodology to overcome shortcomings.

Seventeen sources expressed doubt in the original finding based on how '*surprising*' they perceived it to be. While novel findings can be surprising (see, for example, the account in Case 25), this code is distinct in that respondents clearly mentioned their disbelief in the original findings, which was not necessarily true for novel findings per se. Respondents, were surprised by findings "*that were different from what one would expect from general experience*" (Case 8), that is, they were "*unexpected/counterintuitive*" (Case 32). Replicating the surprising findings was a way to "*ensure that the conclusion was right*" (Case 37). It appeared that some participants were more inclined to replicate studies for which they did not believe in the finding. One respondent made this explicit saying: "*in general, I look for papers that I don't believe the findings*" (Case 78). This is in contrast to those who are interested in replicating to build on the original finding.

Doubt could also be due to the statistical evidence appearing weak to the participant (n= 15). This could be due to small sample sizes, weak methodology, large effect size and associated confidence interval, high p-values, weak statistical evidence as measured through Bayes factors, or peculiar statistical analysis. In some cases, concerns about the statistical evidence coincided with concerns about potential questionable research practices (QRPs). Respondents mentioned p-values showing "*peculiar pattern, with many p-values close to the significance threshold*" (Case 35) or that "*the initial statistics were very p-value based (indicating a desire to get a p<0.05)*" (Case 98). Others mentioned "*analytical creativity*" (Case 104) causing doubt. Additional, respondents mentioned no analytical reproducibility, preregistration, or sample size planning, all of which called into question the original finding and motivated (mostly close) replication for the participant.

Failed previous replication attempts further motivated 14 participants to replicate. Respondents mentioned trying to build on the OS, which included an initial replication of the original effect that failed. Consequently, they decided to run a planned replication instead. For example, one participant "*tried to follow up the work [the original authors] did and so first replicated it. Because the replication failed (non-significant results), [the RAs] tried again*" (Case 21). Another respondent shared that they "*tried to build on a new and interesting finding but after several attempts found no effect at all. That is when one of [their] co-authors suggested to go back to the original study and try to replicate that first*" (Case 45).

The lack of replication studies or replications outside the original author's lab similarly caused uncertainty and doubt about the original effect in some participants' minds. The lack of "*internal or external*" replications resulted in the original finding not appearing convincing (e.g., Case 9). Still, only internal replication (i.e., as opposed to external corroboration) could also raise reasonable doubt (e.g., Case 7). Respondents also argued that the lack of previous replication studies made it "*easier for reviewer to see the relevance of a replication*" (Case 13).

Respondents (n=13) also mentioned doubt if the original finding was not in line with the current theory or if the literature provided mixed support for the effect. This was true for older studies, which were not further supported by more recent data or novel studies calling into question the current theory. Respondents mentioned the finding being "*out of line with existing work*" (Case 41) as a motivation to replicate. It seemed that the participants were interested in

verifying the original finding before trying to explain why the effect was not in line with the literature or theory.

Lastly, issues with the original author made some respondents doubtful about the original finding. Respondents expressed doubt if "*the author was ambiguous when [they] asked them for help*" (Case 15) or were not willing to share their data or materials. A few respondents (n = 3) also explicitly cared about the original author's reputation, though another respondent stated that they "*do not care about [. . .] author*" (Case 109). However, for one participant, the reputation of the original author even increased confidence in the original effect "*we knew the original author and found him trustworthy*"(Case 76). Similarly, many participants (n=20) mentioned cooperating with the original authors, which for some was explicitly positive. For example, one participant mentioned that they "*were able to run [their] replication effort thanks to the willingness of the original author to share their data, stimuli, and instructions*" (Case 77).

Methodology Participants (n=77) mentioned several methodological aspects of the OS motivating their decision to replicate, with some (n = 8) making their decision to replicate contingent on specifics of the original method (e.g., "*needed to be carried out with child or adolescent participants*", Case 14). Sample size was the most frequently (n = 26) mentioned concern. Respondents mentioned the original sample being "*rather small*" (Case 9), criticised that the original sample size had not been justified, or expressed their motivation to collect a larger sample. Sample size concerns could be linked to concerns about the effect size of the OS. Respondents specifically mentioned studies with small sample and large effects being in need of replication. Moreover, these concerns were amplified if the study was not preregistered. For example one participant judged that their target finding " *did not seem very credible (small N/large effects sizes/not preregistered)*" (Case 114).

Respondents (n=16) were also concerned with the generalizability of the OS. Generalizability means that RAs examined whether the original finding would extend to different stimuli, settings, or populations. Consequently, generalizability was a frequent concern for replicators, who already had access to a different population than the OS. This code further connected to participants mentioning the demographics of the target population for their replication. For example, one respondent said that their "*replication used very similar methodology, but extended the research question to a different population with greater representation of the clinical symptoms [they were] interested in studying*" (Case 49). It appeared that some respondents found replications especially valuable if they could examine a population different from the OS. One participant made this explicit saying that "*[they] also had the opportunity to collect data from a population demographically different from the original study, increasing the value of the replication*" (Case 72). However, another participant judged it important to use "*a sample as similar as possible*" (Case 85). Notably, most respondents concerned with generalizability and extending the effect self-identified as conducting close replications.

Methodological aspects of the OS could induce doubt in the 'truthfulness' of the original finding. Outdated methods were frequently mentioned (n= 9). In some instances, outdated methods prompted doubt. For example, "*advances i[n] methodological sophistication and quality prompted reconsideration of prior findings that were published using, now, outdated methods*" (Case 30). Other times, outdated methods did not induce doubt but were considered when updating the methodology to fit the current context. For example, "*the statistical analysis we used were updated to reflect advancements in the capabilities of statistical software*" (Case 42) or "*we used updated and better validated measures*" (Case 32).

Respondents were further concerned with potential confounds biasing the original finding. Participants "*chose [. . .] [the] study because [they] thought there was a confound in the experimental design*" (Case 38) and consequently controlled for "*a factor the original authors hadn't*" (Case 94). One explicitly mentioned confound, was experimenter bias. For example, respondents worried about the potential influence from experimenter bias which leads to "*doubt about methodology*" (Case 103) or as another respondent put it: "*[. . .] I was afraid that the original study was suffering*

from experimenter bias" (Case 4). Similarly, this prompted participants to replicate with updated methods.

Respondents (n=5) further mentioned statistical significance as an "*implicit criterion*" (Case 14). Participants were mostly interested in replicating studies for which "*results supported the hypothesis*" (Case 93), though one person explicitly mentioned "*the null result*" (Case 19) as motivating their choice to replicate.

Methodological aspects could also be linked to feasibility considerations. More specifically, some participants (n=9) mentioned that they were interested in replicating simple studies specifically, "*which could be replicated easily and quickly*" (Case 18). This criterion was predominantly applied to student projects.

Infrequently mentioned considerations included the number of trials (n=3), practicing specific statistical analyses (n=2), or replicating the OS with the same sample as previously used (n=1).

(v) Limitations

Results from the survey need to be considered in light of some limitations. First, some participants misunderstood the instructions and answered the items with replications in general in mind instead of the specific replication study that was the basis for us approaching them. This means that some participants reported concerns that were more general and broad. This might account for some discrepancies and some of the variability in the ratings. For example, participants might simultaneously (1) believe that replication should be concerned with generalizability *in principle*; (2) have not considered it a relevant aspect in the decision to conduct their own replication study.

Asking participants to classify their own study as a direct/close or conceptual replication also means that many people will have applied labels according to different criteria or based on different understandings of the concepts of direct/close and conceptual replication. For instance, many participants that conducted their replication study (partly) to extend the original design or to include additional conditions classified their study as a close replication (with extensions). Nonetheless, one could argue that these cases could be classified as conceptual replications. Our results highlight the variability in replication aims and procedures, and the fact that names and definitions are used somewhat interchangeably and vaguely in the literature. In our view, the dichotomous distinction between the two types of replication is not very informative. Defining replication types based on what they might achieve, or going even deeper [50–52] might be a better approach.

(vi) Changes based on Survey

The most frequently reported codes were identified by counting how often themes were mentioned across cases (i.e., how many participants mentioned a code). Counts ranged from 1 to 34 with Mdn = 10. Codes with 10 or more mentions (n=38) were evaluated by the author team. Authors M.-M.P., P.M.I, A.E.v.t.V., and D.v.R. read through the list of frequently mentioned codes, tried to identify connections, linked them back to the preliminary list of considerations, and suggested edits.

M.-M.P. and D.v.R. independently summarized the suggestions and both created a suggestion for a revised version of the list of considerations each. M.-M.P. merged the two suggestions and created a first draft of the revised list. Over the course of three rounds, this draft was further revised by the author team with M.-M.P. summarizing co-authors' feedback between rounds. The intermediate list revisions are detailed in the supplementary material (*List revisions*).

The revised list included 18 items clustered around the six most frequently mentioned themes: (1) interest, (2) doubt, (3) impact, (4) methodology, (5) feasibility, and (6) educational value. These themes partially overlapped with the four themes considered during stage 1, namely uncertainty (here doubt), value/impact, quality (here methodology), and cost/feasibility. Table 4 contains the 18 items (i.e., the rows that have an entry in column "Round 1").

3. Stage 3

(i) Deviations from preregistered plan

After 5 weeks of data collection, data of 32 respondents was downloaded. However, 5 responses were empty, leaving a total of 27 participants. We had initially planned to continue recruitment for three months or until reaching 30 participants. However, in light of the fact that the summer months were coming up, we decided that it was better for the quality of the data to proceed with the Delphi process rather than wait two more months for the last 3 participants to potentially join. We requested permission for this deviation from the editorial office and received approval on June 7th 2022.

The consensus procedure was stopped after three instead of four rounds, even though we did not reach consensus on one item. Specifically, we observed diverse responses with very little movement between rounds despite revisions of the item (Round 1: *Mdn*=6, *IQR* 3, Round 2: *Mdn*=6, *IQR*=3.5, Round 3: *Mdn*=7.5, *IQR*=3). We reasoned that burdening participants with an additional survey round would not lead to consensus on this item. We requested permission for this deviation from the editorial office and received approval on Sep 13th 2022.

(ii) Participants

A total of 63 participants were contacted and invited to participate in the Delphi procedure on 25.04.2022. Additional to the 29 potential participants a priori identified, 34 survey participants indicated interest in participating. We received 27 responses in the first round, and 20 in the second and third round. During the third round, four participants responded twice. We followed up with these participants and included the response, which they identified as most closely reflecting their opinion.[12]

Participants were diverse across career stage, field of expertise, gender, and geographical location. Participants included five PhD candidates[13], three post-doctoral researchers, eleven senior researchers, and one independent researcher. Participants stemmed from various (psychological) fields including psychological methods and statistics, cognitive and experimental psychology, social psychology, clinical and personality psychology, legal psychology, but also philosophy, empirical aesthetics, and (cognitive) neuroscience. Participants identified as men, women, or other. Geographical locations were diverse, but we were unable to recruit participants from South America, Africa, Australia, or the Caribbean or Pacific Islands.

(iii) Results

Overall, three Delphi rounds were conducted. Table 4 summarizes the quantitative results and qualitative changes across the three rounds. Detailed summary reports sent to the participants between rounds can be found on OSF.

During the first round, consensus was established for 12 out of 18 considerations. Based on the preregistered criteria 10 considerations with a median rating of 7 or higher and an IQR of 2 or lower were included, and two considerations with a median lower than 7 and an IQR of 2 or lower were excluded from the final list. No consensus was reached for the remaining six items. Two out of the six items were revised based on the qualitative results.

During the second round, we did not reach consensus for the remaining six items. However, the qualitative input allowed us to revise all items as well as provide some clarifications regarding the aim of the checklist. Specifically, we clarified that the aim of the checklist is to transparently communicate one's rationale for selecting a particular study and not whether a study generally needs to be replicated or not.

During the third round, consensus was established for all but one item. Based on the preregistered criteria three considerations were included, and two considerations were excluded from the final list. No consensus was reached for the remaining item and responses were

[12] An analysis with all responses is presented in the Stage 3 summary report on OSF
[13] One PhD candidate is also a Research fellow

particularly varied ranging from 1 = not at all important to 9 = very important. As a result, this item is not included in the final checklist.

The final checklist included 13 out of 18 items centered around the topics: interest, doubt, impact, methodology, and feasibility. Please consult the supplementary material for the final version of the checklist.

Table 4: Quantitative results of stage 3 checklist development. Included items are highlighted in bold and revisions are indicated in italic.

Item	Round 1		Round 2		Round 3		Decision
	Mdn	IQR	Mdn	IQR	Mdn	IQR	
The relevance of the original study for your current line of research or the field you work in.	7	2					Include
Your involvement in the line of research that the replication target is concerned with (e.g., self-replication, planning to build on the study in the future).	6	3					Revise
The degree of involvement you have in previous or upcoming projects related to the replication target (e.g., self-replication, planning to build on the study in the future).			6	3.5			Revise
Your personal stakes in the replication target's results (e.g., self-replication, financial stakes or other potential conflicts of interest, planning to build on the replication target results in future research, etc.).					7.5	3	No consensus
The current strength of evidence in favour of the original claim (e.g., a high/low Bayes factor, a wide/narrow confidence interval, a high/low p-value).	7	1					Include
Your personal belief about the truthfulness of the original claim (e.g., consensus in findings, replication attempts).	5	2					Exclude
Your expectations about whether the original claim would replicate or not.	5	2					Exclude
The importance of the original study for research (e.g., often/rarely cited, under/over-studied, published in high/low impact journal).	7	1.5					Include
The theoretical relevance of the original claim.	8	2					Include
Implications of the original claim (e.g., for practice, policy or clinical work).	8	2					Include
The clarity and replicability of the original protocol (e.g., completeness and clarity of the methodological description, accessibility of the materials).	6	4					Re-evaluate
-			4	4.25			Revise
The (un)clarity and (un)replicability of the original protocol (e.g., completeness and clarity of the methodological description, accessibility of the materials).					7.5	2	Include
The sample size of the original study (too small or too large).	7	2					Include
Flaws of the original design (e.g., in- an exclusion criteria, potential confounds).	8	1.5					Include
Operationalization of the original study's measures (e.g., validity, reliability, and bias).	7	3					Re-evaluate
-			7	2.25			Revise
Operationalization of the original study's measures (e.g., validity, reliability, and bias) and how this impacts the credibility of the original study.					7	1.25	Include
Concerns that questionable research practices have been employed (e.g., presence/absence of preregistration, potential of p-hacking or HARKing).	7	2					Include
Generalizability of the original finding (e.g., cultural and temporal context, representativeness of the sample).	7	2					Include
The resources available to you for replicating the original study (e.g., funding, time, equipment, study materials, or data).	8	2					Include
The adaptability of the original study design (e.g., mode of data collection, whether the study can be translated into other languages, contexts).	6	2.5					Re-evaluate
-			6	2.25			Revise

Note: Re-evaluate means that participants received qualitative feedback and were asked to rate the same item again.

Table 4 continued.

Item	Round 1 Mdn	Round 1 IQR	Round 2 Mdn	Round 2 IQR	Round 3 Mdn	Round 3 IQR	Decision
The adaptability of the original study design (e.g., whether data is collected online or on-site, whether the study can be translated into other languages or applied to different contexts, etc.).					6.5	1	Exclude
Your previous experience and expertise with regards to the original study.	5	4					Revise
You (i.e, all replicating authors) previous experience and expertise with regards to the original study.			5.5	3			Revise
Your (i.e., the replicating team as a whole) presence or absence of previous experience or expertise on the original study as a practical concern.					7	2	Include
Educational value of conducting the replication study (e.g., for a thesis or student project).	5	3.5					Re-evaluate
-			3	4			Re-evaluate
-					5	1.5	Exclude

Note: Re-evaluate means that participants received qualitative feedback and were asked to rate the same item again.

4. Discussion

(a) Checklist for transparent reporting of replication target selection

Our goal was to develop a checklist for transparent and systematic reporting of the process of replication target selection. Our consensus-based checklist was designed to guide social scientists through the process of selecting a replication target study, and give them a framework for reporting their decisions and justifications. Checklist item selection was informed by two sources: 1) scientists' practices, revealed by a qualitative analysis of survey data, and 2) expert opinions, explored through a Delphi panel discussion.

Importantly, this checklist covers reasons why a study *was actually selected*, not a list of reasons why a study *ought to be selected*. That is, rather than reporting whether a study needs to be replicated in general, the checklist aims to transparently communicate one's rationale for selecting a particular study. We initially planned to create a list of items which ought to be *ideally* considered when selecting a replication target (see also the specification in Figure 1). However, the survey illustrated the variety of potential reasons to select a replication target and underscored the need for transparency, more so than validity of the items. For example, while some might consider it invalid to replicate a study because it was easy to do (the relevant know-how was already present in the team), this reasoning frequently informed replication target selection in practice. Consequently, we moved away from what to ideally *consider* towards what to ideally *report*. Specifically, we asked our Delphi participants to consider that if the researcher used a consideration as a ground for replicating (irrespective of their personal assessment of the legitimacy of that reason), was it important for that reason to be explicitly communicated? The checklist can either be used to compare several targets for replication in an attempt to identify and justify the chosen replication target, or to report the justification for having chosen a specific replication target after the fact.

We argue that our checklist will enable evaluation of future decisions to replicate and aid discussion about how resources are allocated, and which studies ought to be prioritized. Our checklist will also be useful to assist replicating researchers in explicate their decision process as they prepare their study protocol. The checklist could also be used to evaluate funding applications for replication studies. For the purpose of justification and decision-making, we advise researchers to complete this list before the start of a replication project. For the purpose of documentation, researchers might complete this list at a later time point. However, we caution that hindsight bias might affect the accuracy of the information if the checklist is filled out after the project is complete.

(b) The guiding principles of replication target selection

Checklist items are grouped according to five themes that we constructed from the survey data: (1) interest, (2) doubt, (3) impact, (4) methodology, and (5) feasibility. This theme structure is validated by similar findings in the literature, such as those of Isager and colleagues [16,53]. Reviewing 68 self-reported justifications for replication target selection, Isager [53] identified four factors guiding replication target selection: (1) uncertainty, (2) value/impact, (3) quality, and (4) feasibility. While we initially adopted the structure proposed by Isager and colleagues [16], it was abandoned during Stage 1 as we were unable to clearly group items to one theme or another. Seeing that we independently reconstructed these themes during the present survey lends further credibility to them being the guiding principles of replication target selection. Note however, that we cannot exclude the possibility that we surveyed some authors whose replications were also reviewed by Isager [53]. A quick search demonstrated that some of the potential survey participants we identified were also listed in the Curated Replications Table on `curatescience.org`. Nonetheless, the present survey included more potential participants of replications published after 2017[14] than before, so potential overlap should be minimal. Moreover, in the present project the qualitative analysis was performed by M.-M.P. and S.M.F., without the input from P.M.I.

We identified four stable principles that likely underpin replication target selection: *doubt/uncertainty*, *impact/value*, *methodology/quality* and *feasibility/cost*. These are complex constructs, whose meaning and interpretation include several factors, as illustrated by the nested structure of the checklist for transparent replication target selection. Still, researchers looking to strategically choose which study to replicate can use these themes to guide their decision-making process. For any study considered for replication, researchers might ask: (1) Is there reason to doubt the findings? (2) Is the topic important? (3) Are the methods capable of saying something meaningful about the topic under study? (4) Is it feasible to replicate the study in a way that will meaningfully reduce doubt about the findings? We argue that since all four factors interact in generating replication value [for a formal definition of replication value see 16] the answer to all four questions above should be "yes" before a replication is undertaken.

While these principles are a good starting point, each researcher still needs to decide what it is that makes a claim doubtful, have impact, speak to the underlying research question(s), and its methods feasible to be attempted again. The checklist we constructed yields a transparent strategy to select a replication target and guides researchers through these four principles, while providing pointers on how to assess them to avoid arbitrary decisions. This might counteract one notable if unwelcome feature of the replication movement in the 2010's - the contentious atmosphere in channels such as society publications and social media [54]. Specifically, some proponents of replication have taken a maliciously gleeful tone in greeting non-replications, while replication efforts have conversely been disparaged as motivated by hostility and destruction. While explicitly hostile motives were unlikely to emerge from our method based on self-generated explanations of replication research, the controversy does point to need to clarify the prescriptive grounds for the decision to replicate. For example, doubts based only on hunches or suspicions may cover up inadmissible biases and it is better to base doubt-based selection on clearly expressed arguments from prior theory or evidence.

The checklist, when used for transparent reporting, can further shed light on the weight placed on each factor by the RAs. It does not prescribe how to judge each of the items allowing for subjectivity and variability between researchers and contexts to enter the process of replication target selection. This might help to develop individualized strategies for deciding what to replicate, each serving different interpretations of what "uncertainty", "value", "quality", "cost" – and hence, "replication value" – means. This in turn could inform the definition and quantification of replication values.

We identified two additional guiding principles, which were comparably less stable: *interest*, and *educational value*. Personal interest, also mentioned by Isager [53], was frequently mentioned

[14] the cutoff time for [53]

as an internal motivating factor during the survey. However, expert opinions differed whether this item *should* play a role in the decision-making process. Respondents agreed that the relevance of the original study for the RA's line of research plays a role in replication target selection and ought to be communicated. However, they were conflicted about the nature and importance of the RAs' involvement in the original study. Some argued that the RAs should not "*need to have a personal investment in the outcome/line of research*" or considered personal investment as harmful as "*it is also important that the research is designed and conducted impartially*". Others argued that "*scientific and societal stakes should supersede any personal stakes*". However, it appears that personal interest plays a role in replication target selection in practice. Indeed, we cannot assume that scientific stakes will be at odds with personal stakes in cases where personal interest (partly) motivates a replication target selection, especially given that many people's personal interests involved the belief that the OS was interesting, important and worth reinforcing with replication. Moreover, replication context aside, personal interest is a common reason for a researcher to select any given research topic [55], and, some of us argue, a valid one. Should we constrain replication target selection such that personal interest is not part of the decision-making process? We argue that providing a transparent report of the decision-making process in replication target selection largely mitigates the potential risks of allowing personal interest as a motivation for replication.

Some participants reported that in their experience "*self-replication was indeed a strong and primary motivation*" or suggested that "*it is important that researchers are also invested in replicating their own work*". As a result it "*would be important to disclose conflict of interest [. . .] as it might point to bias*". Ultimately, no consensus was reached for this particular item. Controversy may nonetheless be a good reason for RAs to report their personal interest in a topic transparently.

Additionally, we observed replication attempts being conducted for educational purposes, either as seminar classes, theses, or joined research projects. This is in line with the increasing calls to use replication studies as didactic tools [see for example 56]. However, during the Delphi process experts perceived educational value as a secondary benefit of replication studies rather than a guiding principle of what to replicate. In other words, replication was perceived to have educational benefits regardless of which study is replicated.

(i) Close and conceptual replications

The checklist for transparent reporting of replication target selection can be used for different types of replications. Our survey results suggested few differences in considerations between close and conceptual replications[15]. Specifically, concerns regarding generalizability were more frequently mentioned for conceptual replications, whereas motivation to avoid false-positives was more frequently mentioned for close replications. This difference is in line with the functionality of conceptual and close replications identified by Schmidt [57] and described by Zwaan et al. [58] as: "*Direct replications are useful for reducing false positives (i.e., claims that a specific effect exists when it was originally a chance occurrence or fluke), whereas conceptual replications provide information about the generalizability of inferences across different ways of operationally defined constructs and across different populations*" [p. 4, 58].

However, we noticed many instances of a discrepancy between the label participants self-selected for their replication and the label we would have defined based on their description of the purpose of their replication. For example, respondents of close replications aimed to investigate "*a[n] specific effect with a new paradigm*" (Case 1) or "*the boundary conditions of phenomena*" (Case 12), aims that are traditionally assigned to conceptual replication [58]. Other times, following the original research protocol but changing small aspects such as outdated measures (e.g., Case 32: "*[we changed] nothing about the procedure but we used updated and better validated measures*") or imprecise measures (Case 111: "*we used a different measure than originally used that gave us a more precise measure of . . .* ") resulted in respondents labelling their replication as conceptual.

[15]This might partly be due to the small proportion of RAs identifying their replication as conceptual (n=17) versus close (n=94), making the summary statistics for this group less stable.

Conceptual and close replications are thought to be the two ends of a continuum [59] and we indeed observed cases which situated themselves along the continuum but not at either end (e.g., "*It was somewhere in the middle of direct and conceptual*", Case 25). Other respondents described their replication as mixed (e.g., "*We used both*" Case 30) or as close with a conceptual extension (e.g., "*We combined direct replication [. . .] and a conceptual extension [. . .].*", Case 50). We did not, however, observe clear cut-off points, as for example proposed by LeBel and colleagues [60] on the continuum between close and conceptual. Minor changes (to for example the target sample or measures) were sometimes classified as close replications and other times prompted the respondent to identify their replication as conceptual.

Overall, it appears that the distinction between close and conceptual replications in practice is fuzzy at best. At times, this led to questionable scientific conduct. For example, one participant shared that while they conducted a close replication (only varying data analysis), reviewers required them to change the classification to a conceptual replication. The respondent speculated that this might have been a response "*to ease the shock of negative evidence*" (Case 66). Based on our data, we argue that the distinction between close and conceptual replication to be more of a theoretical than practical distinction. This possibility is given weight by the observation that distinctions between different kinds of replication vary widely in the literature [50,61]. The ambiguity in framing does not reflect the variety in kinds or replications in practice.

(c) Limitations

Our results are limited by arbitrary consensus determination, that is, when do we know that consensus has been reached? This limitation is inherent to Delphi procedures [see for example 62]. There is no agreed-upon threshold for consensus in the literature and the present use of a median of 7 with an IQR of 2 was based on previous consensus-based checklist developments [specifically 38]. However, in two instances (e.g., items regarding adaptability and pragmatism) responses were not as stable as anticipated, and whether or not an item was included hinged on the selection of responses. More precisely, the decision to in- or exclude the item changed based on which of the double responses were included in the analysis (see also the summary report for the third Delphi round). In all other instances, we observed stable ratings regardless of which responses were included. We nonetheless caution readers to perceive our checklist as complete and encourage researchers, funding agencies, and other research bodies to provide feedback and recommendations. Moreover, they might want to consider adapting the checklist to their needs.

Additionally, we noted that more than half of our survey respondents came from cognitive and experimental and social psychology, potentially limiting the generalizability of our survey results. One potential explanation might be that the replication crisis in psychology rooted in social and experimental psychology [e.g., 4,5] and calls for replications appeared earlier in social psychology making the practice more wide-spread in these sub-fields. Nonetheless, as our Delphi participants varied in their expertise, and as many of the concepts yielded by our analyses are applicable outside of these fields, we believe our results to generalize to most branches of social science.

(d) Conclusion

Replication target selection appears to be guided by four principal factors: (1) "doubt/uncertainty", (2) "impact/value", (3) "methodology/quality" and (4) "feasibility/cost". Replication target selection is multi-faceted and strategies for deciding what to replicate might depend on the subjective interpretation of the guiding principles. Our checklist for transparent reporting of replication target selection offers one conceptualization of these factors and prompts researchers to consider these themes when selecting a replication target. Moreover, it facilitates conversation about which studies to select for replication by providing a unified framework for how to approach and communicate such decisions.

Authors' Contributions. Conceptualization: M.-M.P., S.M.F., P.M.I., A.E.v.t.V., and D.v.R.; Data curation: M.-M.P.; Formal analysis: M.-M.P. and D.v.R.; Funding acquisition: D.v.R.; Investigation: M.-M.P., T.A., S.N.C., T.D., R.G.-S., S.G., T.H., M.J., T.J.L., D.B.M., R. Peels, R. Pendrous, S.S., J.M.S., E.S., U.S.T., M.A.V., J.M.W., N.Y., and R.A.Z. Methodology: M.-M.P., S.M.F., P.M.I., A.E.v.t.V., and D.v.R.; Project administration: M.-M.P.; Resources: M.-M.P.; Supervision: D.v.R.; Visualization: M.-M.P.; Writing - original draft: M.-M.P.; Writing - review & editing: S.M.F., P.M.I., A.E.v.t.V., and D.v.R.

Acknowledgements. We would like to thank Balazs Aczel and Barnabas Szaszi for their input regarding the Delphi procedure. Additional thanks to Joyce M. Hoek for her input, guidance, and reassurance on qualitative data analysis and Ymkje Anna de Vries for her regular input and supervision on the project.

Funding. M.-M.P. and D.v.R. were supported by an NWO Vidi grant to D.v.R. (016.Vidi.188.001)

References

1. Pashler H, Wagenmakers E. 2012 Editor's Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence?. *Perspectives on Psychological Science* **7**, 528–530. Publisher: SAGE Publications Inc.
2. Baker M. 2016 Dutch agency launches first grants programme dedicated to replication. ISSN: 14764687 Publication Title: Nature.
3. Ioannidis JPA. 2005 Why Most Published Research Findings Are False. *PLoS Medicine* **2**, e124. Publisher: Springer International Publishing.
4. Levelt Committee, Noort Committee, Drenth Committee. 2012 Flawed science: The fraudulent research practices of social psychologist Diederik Stapel. Technical report.
5. Wagenmakers EJ, Wetzels R, Borsboom D, van der Maas HL, Kievit RA. 2012 An Agenda for Purely Confirmatory Research. *Perspectives on Psychological Science* **7**, 632–638. Publisher: Perspect Psychol Sci.
6. Simmons JP, Nelson LD, Simonsohn U. 2011 False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* **22**, 1359–1366. Publisher: SAGE Publications Inc.
7. Fiedler K. 2011 Voodoo Correlations Are Everywhere-Not Only in Neuroscience.. *Perspectives on psychological science : a journal of the Association for Psychological Science* **6**, 163–71. Publisher: SAGE Publications Inc.
8. Makel MC, Plucker JA, Hegarty B. 2012 Replications in psychology research: How often do they really occur?. *Perspectives on Psychological Science* **7**, 537–542. Publisher: SAGE PublicationsSage CA: Los Angeles, CA.
9. Open Science Collaboration. 2015 Estimating the reproducibility of psychological science. *Science* **349**, aac4716. Publisher: American Association for the Advancement of Science.
10. Isager PM. 2019 Quantifying Replication Value: A guide in the decision of what to replicate. .
11. McNeeley S, Warner JJ. 2015 Replication in criminology: A necessary practice. *European Journal of Criminology* **12**, 581–597. Publisher: SAGE Publications.
12. Errington TM, Iorns E, Gunn W, Tan FE, Lomax J, Nosek BA. 2014 An open investigation of the reproducibility of cancer biology research. *eLife* **3**, e04333.
13. Cook SC, Schwartz AC, Kaslow NJ. 2017 Evidence-Based Psychotherapy: Advantages and Challenges. *Neurotherapeutics* **14**, 537–545.
14. Institure of Educational Sciences. 2019 IES FY 2020 Request for Applications Research Grants Focused on Systematic Replication CFDA Number: 84.305R. Technical report.
15. QUEST The QUEST Replication Call Background. .
16. Isager PM, van Aert RCM, Bahník S, Brandt MJ, DeSoto KA, Giner-Sorolla R, Krueger JI, Perugini M, Ropovik I, van't Veer AE, Vranka M, Lakens D. 2021 Deciding what to replicate: A decision model for replication study selection under resource and knowledge constraints. *Psychological Methods* pp. No Pagination Specified–No Pagination Specified. Place: US Publisher: American Psychological Association.
17. Laws KR. 2016 Psychology, replication & beyond. *BMC Psychology* **4**, 30–30.
18. Field SM, Hoekstra R, Bringmann L, van Ravenzwaaij D. 2019 When and Why to Replicate: As Easy as 1, 2, 3?. *Collabra: Psychology* **5**, 46–46. Publisher: PsyArXiv.

19. Pittelkow MM, Hoekstra R, Karsten J, van Ravenzwaaij D. 2021 Replication target selection in clinical psychology: A Bayesian and qualitative reevaluation.. *Clinical Psychology: Science and Practice* **28**, 210–221.
20. Coles NA, Tiokhin L, Scheel AM, Isager PM, Lakens D. 2018 The costs and benefits of replication studies. *Behavioral and Brain Sciences* **41**, e124. Publisher: Cambridge University Press.
21. Hardwicke TE, Tessler MH, Peloquin BN, Frank MC. 2018 A Bayesian decision-making framework for replication. *Behavioral and Brain Sciences* **41**, e132–e132. Publisher: Cambridge University Press.
22. Kuehberger A, Schulte-Mecklenbeck M. 2018 Selecting target papers for replication. *Behavioral and Brain Sciences* **41**, e139. Publisher: Cambridge University Press.
23. Nuijten M. 2021 Efficient Scientific Self-Correction in Times of Crisis. *The New Common* p. 161. Publisher: Nature Publishing Group.
24. Murphy J, Mesquida C, Caldwell AR, Earp BD, Warne J. 2021 Selection Protocol for Replication in Sports and Exercise Science. type: article.
25. Waggoner J, Carline JD, Durning SJ. 2016 Is there a consensus on consensus methodology? Descriptions and recommendations for future consensus research. *Academic Medicine* **91**, 663–668. Publisher: Wolters Kluwer.
26. McKenna HP. 1994 The Delphi technique: a worthwhile research approach for nursing?. *Journal of Advanced Nursing* **19**, 1221–1225. Publisher: John Wiley & Sons, Ltd.
27. LeBel EP, McCarthy RJ, Earp BD, Elson M, Vanpaemel W. 2018 A Unified Framework to Quantify the Credibility of Scientific Findings. *Advances in Methods and Practices in Psychological Science* **1**, 389–402. Publisher: SAGE Publications Inc.
28. HÃijffmeier J, Mazei J, Schultze T. 2016 Reconceptualizing replication as a sequence of different studies: A replication typology. *Journal of Experimental Social Psychology* **66**, 81–92. Publisher: Academic Press.
29. Muradchanian J, Hoekstra R, Kiers H, van Ravenzwaaij D. 2021 How best to quantify replication success? A simulation study on the comparison of replication success metrics. *Royal Society Open Science* **8**. Publisher: The Royal Society.
30. Field SM, Wagenmakers EJ, Newell BR, Zeelenberg R, van Ravenzwaaij D. 2016 Two Bayesian tests of the GLOMOsys Model. *Journal of experimental psychology. General* **145**, e81–e95. Publisher: J Exp Psychol Gen.
31. van Ravenzwaaij D, Boekel W, Forstmann BU, Ratcliff R, Wagenmakers EJ. 2014 Action Video Games Do Not Improve the Speed of Information Processing in Simple Perceptual Tasks. *Journal of experimental psychology. General* **143**, 1794. Publisher: NIH Public Access.
32. Brandt MJ, IJzerman H, Dijksterhuis A, Farach FJ, Geller J, Giner-Sorolla R, Grange JA, Perugini M, Spies JR, van 't Veer A. 2014 The Replication Recipe: What makes for a convincing replication?. *Journal of Experimental Social Psychology* **50**, 217–224. Publisher: Academic Press.
33. Collaboration OS. 2012 An Open, Large-Scale, Collaborative Effort to Estimate the Reproducibility of Psychological Science:. *https://doi.org/10.1177/1745691612462588* **7**, 657–660. Publisher: SAGE PublicationsSage CA: Los Angeles, CA.
34. Bouwmeester S, Verkoeijen PPJL, Aczel B, Barbosa F, Bègue L, Brañas Garza P, Chmura TGH, Cornelissen G, Døssing FS, Espín AM, Evans AM, Ferreira-Santos F, Fiedler S, Flegr J, Ghaffari M, Glöckner A, Goeschl T, Guo L, Hauser OP, Hernan-Gonzalez R, Herrero A, Horne Z, Houdek P, Johannesson M, Koppel L, Kujal P, Laine T, Lohse J, Martins EC, Mauro C, Mischkowski D, Mukherjee S, Myrseth KOR, Navarro-Martínez D, Neal TMS, Novakova J, Pagà R, Paiva TO, Palfi B, Piovesan M, Rahal RM, Salomon E, Srinivasan N, Srivastava A, Szaszi B, Szollosi A, Thor KØ, Tinghög G, Trueblood JS, Bavel JJV, van't Veer AE, Västfjäll D, Warner M, Wengström E, Wills J, Wollbrant CE. 2017 Registered Replication Report: Rand, Greene, and Nowak (2012):. *https://doi.org/10.1177/1745691617693624* **12**, 527–542. Publisher: SAGE PublicationsSage CA: Los Angeles, CA.
35. Klein RA, Vianello M, Hasselman F, Adams BG, Reginald B. Adams J, Alper S, Aveyard M, Axt JR, Babalola MT, Bahník S, Batra R, Berkics M, Bernstein MJ, Berry DR, Bialobrzeska

O, Binan ED, Bocian K, Brandt MJ, Busching R, Rédei AC, Cai H, Cambier F, Cantarero K, Carmichael CL, Ceric F, Chandler J, Chang JH, Chatard A, Chen EE, Cheong W, Cicero DC, Coen S, Coleman JA, Collisson B, Conway MA, Corker KS, Curran PG, Cushman F, Dagona ZK, Dalgar I, Rosa AD, Davis WE, de Bruijn M, Schutter LD, Devos T, de Vries M, Doğulu C, Dozo N, Dukes KN, Dunham Y, Durrheim K, Ebersole CR, Edlund JE, Eller A, English AS, Finck C, Frankowska N, Freyre MA, Friedman M, Galliani EM, Gandi JC, Ghoshal T, Giessner SR, Gill T, Gnambs T, Gómez A, González R, Graham J, Grahe JE, Grahek I, Green EGT, Hai K, Haigh M, Haines EL, Hall MP, Heffernan ME, Hicks JA, Houdek P, Huntsinger JR, Huynh HP, IJzerman H, Inbar Y, Innes-Ker ÃH, Jiménez-Leal W, John MS, Joy-Gaba JA, Kamiloğlu RG, Kappes HB, Karabati S, Karick H, Keller VN, Kende A, Kervyn N, Knežević G, Kovacs C, Krueger LE, Kurapov G, Kurtz J, Lakens D, Lazarević LB, Levitan CA, Neil A. Lewis J, Lins S, Lipsey NP, Losee JE, Maassen E, Maitner AT, Malingumu W, Mallett RK, Marotta SA, Mededović J, Mena-Pacheco F, Milfont TL, Morris WL, Murphy SC, Myachykov A, Neave N, Neijenhuijs K, Nelson AJ, Neto F, Nichols AL, Ocampo A, O'Donnell SL, Oikawa H, Oikawa M, Ong E, Orosz G, Osowiecka M, Packard G, Pérez-Sánchez R, Petrović B, Pilati R, Pinter B, Podesta L, Pogge G, Pollmann MMH, Rutchick AM, Saavedra P, Saeri AK, Salomon E, Schmidt K, Schönbrodt FD, Sekerdej MB, Sirlopú D, Skorinko JLM, Smith MA, Smith-Castro V, Smolders KCHJ, Sobkow A, Sowden W, Spachtholz P, Srivastava M, Steiner TG, Stouten J, Street CNH, Sundfelt OK, Szeto S, Szumowska E, Tang ACW, Tanzer N, Tear MJ, Theriault J, Thomae M, Torres D, Traczyk J, Tybur JM, Ujhelyi A, van Aert RCM, van Assen MALM, van der Hulst M, van Lange PAM, van't Veer AE, Echeverría AV, Vaughn LA, Vázquez A, Vega LD, Verniers C, Verschoor M, Voermans IPJ, Vranka MA, Welch C, Wichman AL, Williams LA, Wood M, Woodzicka JA, Wronska MK, Young L, Zelenski JM, Zhijia Z, Nosek BA. 2018 Many Labs 2: Investigating Variation in Replicability Across Samples and Settings:. *https://doi.org/10.1177/2515245918810225* **1**, 443–490. Publisher: SAGE PublicationsSage CA: Los Angeles, CA.

36. Field SM, Wagenmakers EJ, Kiers HAL, Hoekstra R, Ernst AF, van Ravenzwaaij D. 2020 The effect of preregistration on trust in empirical research findings: results of a registered report. *Royal Society Open Science* **7**, 181351. Publisher: Royal Society.
37. Alister M, Vickers-Jones R, Sewell DK, Ballard T. 2021 How Do We Choose Our Giants? Perceptions of Replicability in Psychological Science:. *Advances in Methods and Practices in Psychological Science* **4**. Publisher: SAGE PublicationsSage CA: Los Angeles, CA.
38. Aczel B, Szaszi B, Sarafoglou A, Kekecs Z, Kucharský , Benjamin D, Chambers CD, Fisher A, Gelman A, Gernsbacher MA, Ioannidis JP, Johnson E, Jonas K, Kousta S, Lilienfeld SO, Lindsay DS, Morey CC, Munafò MR, Newell BR, Pashler H, Shanks DR, Simons DJ, Wicherts JM, Albarracin D, Anderson ND, Antonakis J, Arkes HR, Back MD, Banks GC, Beevers C, Bennett AA, Bleidorn W, Boyer TW, Cacciari C, Carter AS, Cesario J, Clifton C, Conroy RM, Cortese M, Cosci F, Cowan N, Crawford J, Crone EA, Curtin J, Engle R, Farrell S, Fearon P, Fichman M, Frankenhuis W, Freund AM, Gaskell MG, Giner-Sorolla R, Green DP, Greene RL, Harlow LL, de la Guardia FH, Isaacowitz D, Kolodner J, Lieberman D, Logan GD, Mendes WB, Moersdorf L, Nyhan B, Pollack J, Sullivan C, Vazire S, Wagenmakers EJ. 2020 A consensus-based transparency checklist. *Nature Human Behaviour* **4**, 4–6.
39. Braun V, Clarke V. 2006 Using thematic analysis in psychology. *Qualitative Research in Psychology* **3**, 77–101.
40. Field SM, Ravenzwaaij Dv, Pittelkow MM, Hoek JM, Derksen M. 2021 Qualitative Open Science - Pain Points and Perspectives. .
41. Miles MB, Huberman AM. 1994 *Qualitative data analysis: An expanded sourcebook*. sage.
42. Syed M, Nelson SC. 2015 Guidelines for Establishing Reliability When Coding Narrative Data. *Emerging Adulthood* **3**, 375–387. Publisher: SAGE Publications Inc.
43. Birko S, Dove ES, Özdemir V. 2015 Evaluation of Nine Consensus Indices in Delphi Foresight Research and Their Dependency on Delphi Survey Characteristics: A Simulation Study and Debate on Delphi Design and Interpretation. *PLOS ONE* **10**, e0135162. Publisher: Public Library of Science.

44. Noy C. 2008 Sampling Knowledge: The Hermeneutics of Snowball Sampling in Qualitative Research. *International Journal of Social Research Methodology* **11**, 327–344.
45. Venette S. 2013 What is snowball sampling?. .
46. Jorm AF. 2015 Using the Delphi expert consensus method in mental health research. *Australian & New Zealand Journal of Psychiatry* **49**, 887–897.
47. Murphy MK, Black NA, Lamping DL, McKee CM, Sanderson CF, Askham J, Marteau T. 1998 Consensus development methods, and their use in clinical guideline development.. *Health technology assessment (Winchester, England)* **2**, i–88. Publisher: NIHR Journals Library.
48. Gargon E, Crew R, Burnside G, Williamson PR. 2019 Higher number of items associated with significantly lower response rates in COS Delphi surveys. *Journal of Clinical Epidemiology* **108**, 110–120. Publisher: Elsevier USA.
49. Braun V, Clarke V. 2022 *Thematic analysis: a practical guide*. Los Angeles: SAGE.
50. Clemens MA. 2017 The meaning of failed replications: A review and proposal. *Journal of Economic Surveys* **31**, 326–342.
51. Zwaan RA, Etz A, Lucas RE, Donnellan MB. 2018 Improving social and behavioral science by making replication mainstream: A response to commentaries. *The Behavioral and Brain Sciences* **41**, e157–e157. Publisher: NLM (Medline).
52. Gómez OS, Juristo N, Vegas S. 2014 Understanding replication of experiments in software engineering: A classification. *Information and Software Technology* **56**, 1033–1048.
53. Isager PM. 2018 What to Replicate? Justifications of study choice from 85 replication studies.. Technical report Zenodo.
54. Derksen M, Field S. 2022 The Tone Debate: Knowledge, Self, and Social Order. *Review of General Psychology* **26**, 172–183. Publisher: SAGE Publications Inc.
55. Field SM, Derksen M. 2021 Experimenter as automaton; experimenter as human: exploring the position of the researcher in scientific research. *European Journal for Philosophy of Science* **11**, 1–21.
56. Bauer G, Breznau N, Gereke J, HÃűffler JH, Janz N, Rahal RM, Rennstich JK, SoinÃl' H. 2022 Teaching Constructive Replications in the Social Sciences. .
57. Schmidt S. 2009 Shall we Really do it Again? The Powerful Concept of Replication is Neglected in the Social Sciences. *Review of General Psychology* **13**, 90–100. Publisher: SAGE PublicationsSage CA: Los Angeles, CA.
58. Zwaan RA, Etz A, Lucas RE, Donnellan MB. 2018 Making replication mainstream. *Behavioral and Brain Sciences* **41**, e120. Publisher: Cambridge University Press.
59. Soderberg CK, Errington TM. 2019 Replications and the Social and Behavioral Sciences. In *Advanced Research Methods for the Social and Behavioral Sciences* , p. 229. Cambridge University Press. Publisher: Cambridge University Press.
60. LeBel EP, Berger D, Campbell L, Loving TJ. 2017 Falsifiability is not optional. *Journal of Personality and Social Psychology* **113**, 254–261. Place: US Publisher: American Psychological Association.
61. Schmidt S. 2009 Shall We Really Do It Again? The Powerful Concept of Replication Is Neglected in the Social Sciences. *Review of General Psychology* **13**, 90–100.
62. Wuestefeld A, Fuermaier ABM, Bernardo-Filho M, SÃą-Caputo DdCd, Rittweger J, Schoenau E, Stark C, Marin PJ, Seixas A, Judex S, Taiar R, Nyakas C, Zee EAvd, Heuvelen MJGv, Tucha O. 2020 Towards reporting guidelines of research using whole-body vibration as training or treatment regimen in human subjectsâĂŤA Delphi consensus study. *PLOS ONE* **15**, e0235905. Publisher: Public Library of Science.